Clustering Digital Ego Networks by Tie Strength: A Scalable, Platform-Independent Method

1st Masoud Fatemi
School of Computing
University of Eastern Finland
Joensuu, Finland
Center for Data Intensive Sciences & App.
Linnaeus University
Växjö, Sweden
masoud.fatemi@uef.fi

2nd Mikko Laitinen
School of Humanities
University of Eastern Finland
Joensuu, Finland
Center for Data Intensive Sciences & App.
Linnaeus University
Växjö, Sweden
mikko.laitinen@uef.fi

3rd Pasi Fränti
School of Computing
University of Eastern Finland
Joensuu, Finland
School of Data Science
Chinese University of Hong Kong
Shenzhen, China
pasi.franti@uef.fi

***Abstract*—This study presents a scalable method to classify online social networks based on tie strength. Utilizing ego networks from Twitter, we applied four measurable features—interaction strength, relative interaction strength, social similarity, and outlier ratio—to cluster over 8,000 networks into four categories: weak, moderately-weak, moderately-strong, and strong ties. Our approach is not platform-dependent and overcomes the limitations of previous methods that relied on fixed thresholds or manual labeling. The results reveal regional and gender-based differences in tie strength patterns: Nordic users tend to form weaker ties, while users in Australia, the UK, and the US are more likely to build stronger-tie networks. Male users dominate across all tie categories, while female and uncategorized users are more common in weaker networks. The findings can support research in online social behavior, content delivery, and information diffusion.**

***Index Terms*—Social network analysis, Ego networks, Weak-ties, Strong-ties, Clustering**

I. Introduction

In social network theory, networks are characterized by tie strength [1]. In loose-knit environments such as big workplaces or urban neighbourhoods weak-ties tend to prevail [1]. On the other hand, in close-knit networks like close family or close friend groups, the majority of ties are typically strong-ties [1]. Analyzing and differentiating ties strength is important in various research areas, including information dissemination [2], [3], innovation diffusion [4], [5], social movements [6], industry structures [7], health studies [8], and statistics [9].

It has been argued that in weak-tie networks, such as acquaintances or distant colleague networks, individuals infrequently interact and have lower emotional intensity [2]. However, such networks are valuable for accessing new information, innovations, and diverse opportunities [2]. On the other hand, in strong-tie networks, which mainly consist of close relationships, there are frequent interactions between individuals with emotional support and a high level of trust [1], [6]. These networks are helpful for personal support and deep collaboration but involve redundant information since individuals are more likely to have the same information [1], [6].

In this paper, we focus on digital networks from social media. In today's world, analyzing online social networks is crucial for understanding social dynamics and how relationships influence information flow, given that much of our daily activities take place online [10]. This knowledge helps predict how information and misinformation spread, identifying key networks and users involved. Understanding tie strength can also improve personalized content delivery and enhance targeted advertising. Furthermore, examining the characteristics of networks based on tie strength provides insights into the statistical differences between strong and weak tie environments.

Extensive research exists on detecting tie strength in social media networks [2], [4], [11], [12]. However, a major gap is that existing studies lack a comprehensive algorithmic approach to accurately distinguish ties and classifying them on a scale from truly weak ties to truly strong ones. Most methods rely on a predefined heuristic or a domain-specific threshold to classify networks as weak or strong ties. Additionally, the absence of a fast and automated labeling process limits the scalability and practical application of these methods in large-scale datasets.

We study clustering online social networks from four locations into weak-tie and strong-tie networks. Our main research questions are as follows: First, how can we cluster digital networks from social media based on tie strength, categorizing them into weak-tie and strong-tie networks? Second, from a statistical perspective, what are the characteristics of these clusters, and how do they relate to strong and weak tie networks?

In this paper, we use four measures studied by Fatemi et al. [13] for clustering, calculated for each datapoint in our network dataset. Each datapoint represents an ego network consisting of a central node (ego), its directly connected neighbors (alters), and the connections between alters. "Fig. 1" displays an example of an ego network. We represent each ego network as a vector in a four-dimensional space utilizing the four measures that we extract from each network.

The first measure we use is interaction strength (IS), which

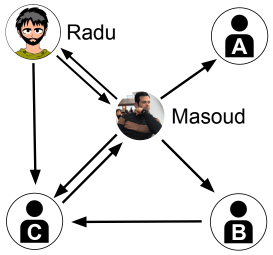

Fig. 1. Masoud's ego network consists of 5 nodes.

represents the weighted interaction frequency among all nodes in an ego network. In Figure 1, for a single directed edge from node *B* to *C*, IS is calculated as:

$$IS_{\overrightarrow{B,C}} = (w_1 \times \text{retweet}) + (w_2 \times \text{quote}) + (w_3 \times \text{reply}) \quad (1)$$

where $\sum w_i = 1$ are the regulation terms, and *retweet*, *quote*, and *reply* represent how many times node *B* retweeted, quoted, and replied to node *C*'s messages. We first calculate IS values for every edge in the network and then compute the average IS value for the entire network. A higher IS value indicates that the network leans more toward strong-tie connections in the weak-to-strong tie spectrum [13]. The second measure, relative interaction strength (RIS), is derived from IS. RIS calculates the proportion of interactions between alters compared to interactions between alters and egos as follows:

$$RIS = \frac{\sum_{a,b \in V:\ a \neq b \neq masoud} IS_{\overrightarrow{a,b}}}{\sum_{a,b \in V:\ a\ or\ b = masoud} IS_{\overrightarrow{a,b}}} \quad (2)$$

where V is the set of nodes and IS is the interaction strength between every pair of nodes such as a and b. Like IS, a higher RIS value signifies a stronger tie network [13].

The third measure we use is social similarity (SS), which is based on the number of shared friends within an ego network as follows:

$$SS = \frac{2}{N \times (N-1)} \sum_{a,b \in V} \frac{|a_{friends} \cap b_{friends}|}{|a_{friends} \cup b_{friends}|}. \quad (3)$$

Here, *N* is the number of nodes in the network, and *V* is the set of nodes. For every pair of nodes in the network, such as *a* and *b*, we calculate the Jaccard similarity value [14] based on their friends' sets. As it is a symmetric similarity value, we exclude the self-similarity values (the main diagonal of the calculated similarity matrix) and normalize the final value by the total number of pairs. Like IS and RIS, networks with higher SS values are considered stronger-tie networks compared to those with lower SS values [13]. The final measure is outlier (OUT), which is calculated based on the structure of the ego network. It represents the percentage of nodes that become completely isolated when the ego node and its links are removed [13]. Unlike the other three measures, lower OUT values indicate a stronger-tie network.

Our findings demonstrate that Nordic social media users build networks differently compared to users from Australia (AU), the United Kingdom (UK), and the United States (US). While users in the UK, AU, and US predominantly form strong-tie networks, Nordic users are more likely to establish weak-tie and moderately-tie networks.

II. Detecting tie strength

In this section, we first review the most relevant work on detecting networks ties strength. Next, we will describe how we compiled our network dataset. Finally, we outline the preprocessing steps applied before clustering the ego networks.

Onnela et al. [11] studied the connection between tie strength and network structure in mobile phone networks. The authors analyzed 18 weeks of call records, covering 20% of a country's population, and compiled an undirected network that included 4.6 million nodes and 7 million edges. In the network, two users were connected if they had at least one reciprocated phone call. The authors measured tie strength by a single number representing the total duration of calls between two users connected by an edge.

Bakshy et al. [2] utilized an undirected network collected from Facebook and measured the tie strength by counting four interaction types: (i) private messages, (ii) public comments on each other's posts, (iii) appearing in the same photo, and (iv) commenting on the same post. Bakshy et al. [2] considered one interaction as the threshold and labeled ties with at least one interaction as strong.

Goel et al. [12] investigated how language changes spread through Twitter. The authors compiled an undirected reply network where nodes were users, and edges were connected if two users replied to each other. In contrast to the network dataset we compiled, Goel et al. [12] preferred the reply network instead of the follower network because of data availability. Goel et al. [12] employed the similarity concept introduced by Adamic and Adar [15] and measured the tie strength based on the normalized mutual friends called embeddedness metric, which gives more weight to low-degree mutual friends.

Del Tredici and Fernández [4] scraped data from 20 Reddit forums to study the strength of ties in emerging and distributing new words on social networks. The authors formed unweighted and undirected monthly networks for each forum where users were nodes and interactions were edges. Del Tredici and Fernández [4] adopted the Onnela et al. [11] approach for measuring relative topological overlap between nodes and computed tie strength ranging from 0 (weakest) to 1 (strongest) only based on the overlap of their adjacent neighborhoods. Nodes' tie strength was also calculated based on their strongest connection.

Feehan et al. [16] utilized a sample of 595 respondents from a survey in Hanoi to form a social network and investigated estimating the size of weak-tie personal networks. Feehan et al. [16] defined four tie groups ranging from weak to strong and asked respondents to evaluate their connections to groups under four tie definitions.

Based on data collected from LinkedIn, Rajkumar et al. [17] measured tie strength by two distinct factors: counting direct messages between users (interaction intensity) and shared friends when a connection was made (mutual connections).

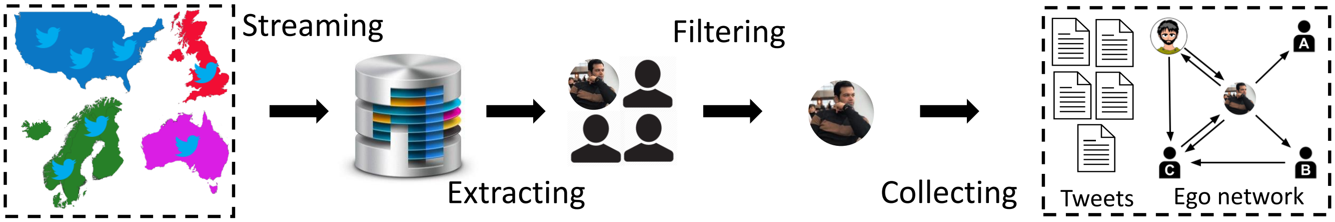

Fig. 2. An overview of the collection process of DSN corpora.

Rajkumar et al. [17] used the median value as the cut-off point in both factors to categorize ties as weak or strong.

A. Data

Twitter (now called X) is the primary data source in this paper. We compiled a geo-located network dataset from Twitter in early 2023, utilizing Python and the now-discontinued Academic Twitter API.

Twitter users may share their locations in their tweet text. However, these types of locations are unreliable and not easy to access. For instance, there are similar names for different places (Paris, France vs. Paris, Texas), or there are common words that can also refer to locations (Orange can be a color, fruit, or city in France). Alternatively, we can use the location section in the Twitter profile to access a user's location. This information is also unreliable because it is a free-text field, allowing users to mention imaginary locations. The last option that is employed is utilizing geo-location-enabled tweets for users who have activated this feature in their profile settings. For these users, standard location information, including the country name and latitude and longitude, will be integrated into every single tweet.

The constructed geo-located dataset is called Digital Social Network Corpora (DSN Corpora) and spans from 2006 to 2022, including 19,345 ego networks comprising 829,608 nodes and 5,066,655 directed edges representing friendship on Twitter. DSN Corpora encompasses four distinct locations: AU, Nordic, UK, and US. "Fig. 2" represents the overall process of data collection.

In step 1, we streamed geo-located tweets from AU, the UK, the US, and the Nordic region. In step 2, we extracted the creators of the tweets collected in the previous step and created a primary candidate list of accounts. As for step 3, to access genuine human networks [5], we cleaned our primary candidate list, filtered out institutional accounts, and verified accounts such as celebrities or policymakers with few friends and millions of followers. We also excluded the very passive and very active accounts [5]. In step 4, for the remaining accounts that passed the filtering process, we collected their ego networks (friend's list and friends' list of friends) and their latest tweets up to 3200 (the limitation imposed by the Twitter API) from all the nodes. Table I represents the basic statistics of the compiled network dataset.

"Fig. 3" top displays the creation dates of the accounts available in DSN corpora and from which data was collected. Users from the UK, AU, and US follow a consistent pattern. However, the Nordic region indicates a different trend, with the number of created accounts increasing as we go back in time until 2016, before beginning to decline.

"Fig. 3" bottom represents the annual tweets collected from each location in the data collection step 4 (see "Fig. 2"). The data spans from 2006, when Twitter was founded, to 2022. Using the Twitter API, we retrieved up to 3,200 of the latest tweets from each Twitter account. This explains the larger volume of data in recent years and the drop in collected tweets as we go back in time for the UK, AU, and US. In contrast, for the Nordic countries, the number of collected tweets decreases over time but increases again between 2017 and 2013 before declining again.

B. Preprocessing

Outlier datapoints are those that simply differ from the majority of the data [18]. Before proceeding to clustering and as a preprocessing step, we cleaned the dataset and also filtered out outliers, resulting in removing some ego networks. We cleaned the dataset first to avoid biased outlier detection in the following filtering step. In other words, removing invalid values ensures that outliers are not detected based on missing or inaccurate data points. Consequently, we set four thresholds based on extracted measures in the cleaning step. We excluded ego networks where IS, RIS, and SS values were zero and those with an OUT value of 100. As a result, 6,324 networks were removed, leaving 13,021 out of 19,345.

After cleaning and removing invalid cases, we applied a filtering step to exclude outlier ego networks by removing those strength measures that deviate the most using some statistical measure such as standard deviation (STD), interquartile range (IQR), or median absolute deviation (MAD). For example, STD-based filtering calculates the mean (μ) and standard deviation and considers a value deviating more than 2×σ from the mean as an outlier.

However, extreme values in the data also make the statistics biased. For this reason, we used the 2T approach by Yang et al. [19], which removes the outliers iteratively. Any of the standard measures (STD, IQR, MAD) can be used within 2T, but with a more conservative threshold (e.g., 3×σ) to remove only the most extreme values. After their removal, the statistics are recalculated, and the process is repeated a few times. This iterative approach makes the outlier removal process more robust and accurate in case of noisy or heavily skewed datasets.

TABLE I
BASIC STATISTICS OF THE COMPILED DSN CORPORA.

Location	Networks	Nodes	Edges	Average degree	Average net size	Tweets
UK	940	54,290	425,280	10.88	58	68M
AU	1,840	201,686	1,299,618	9.87	110	256M
US	2,995	179,484	1,300,649	10.45	60	238M
Nordic	13,570	394,148	2,041,108	6.51	29	190M
Total	**19,345**	**829,608**	**5,066,655**	**7.65**	**43**	**752M**

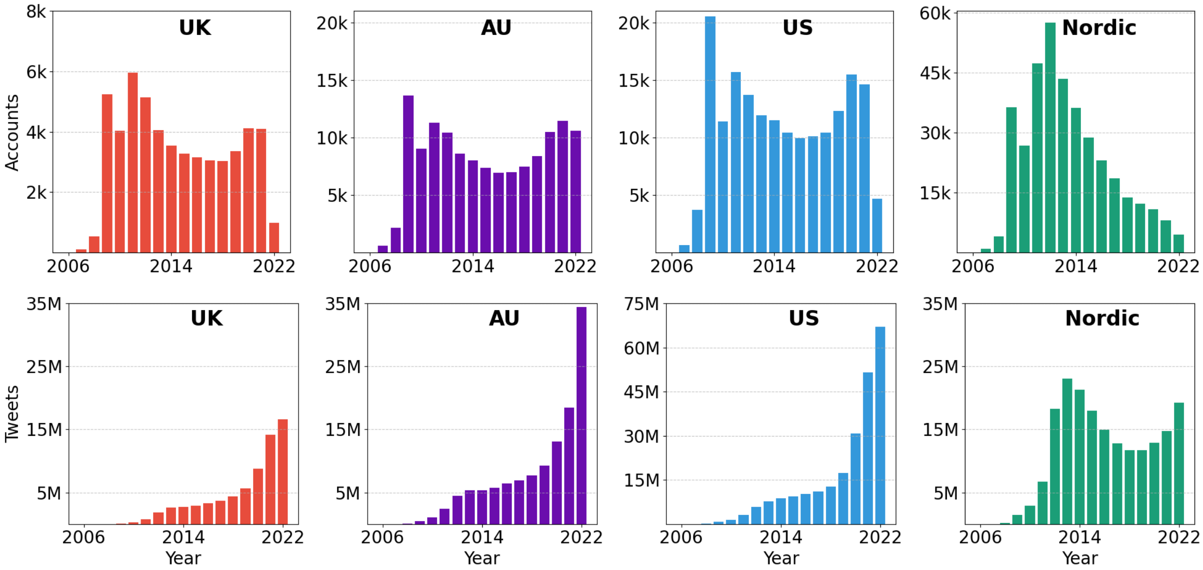

Fig. 3. Top: accounts created date in the dataset per year. Bottom: number of tweets collected per year.

We test 2T with all the three statistics using the parameter selection shown in Table II.

“Fig. 4” (top) illustrates the distribution of one measure, IS, in the raw data across different locations. Due to the heavily tailed distributions, STD-based filtering is not suitable for our data.

IQR-based filtering is more robust to extreme values than STD-based filtering since it relies on percentiles (Q1 and Q3) rather than the mean and standard deviation. However, it can still be influenced by extreme skewness. In highly skewed distributions, see “Fig. 4” top, where data is stretched on one side, the quartiles shift unevenly. This is affecting the filtering process and making the model unsuitable for our data.

As mentioned in Table II, MAD-based filtering is the most robust method compared to the previous two methods when dealing with outlier and highly skewed distributions [20]. Since it is based on median rather than mean and quartiles, it makes it highly resistant to outliers.

Using STD-based filtering, 1,205 networks were excluded from the dataset. Filtering with IQR removed 3,884 networks, while MAD-based filtering excluded 4,848 networks. Due to the robustness of MAD-based filtering against outliers [20], we proceeded with this method, leaving 8,173 networks in the dataset. After cleaning and filtering outliers, there are 383 networks from the UK, 991 from AU, 1,284 from the US, and 5,515 from the Nordic region. “Fig. 4” (bottom) displays the IS measure distributions across these locations after cleaning and applying MAD-based filtering.

III. CLUSTERING

In clustering, the goal is to group a set of objects in a way that objects in the same cluster are more similar to each other than to those in other clusters [21]. In other words, we have a set of data points as $X = \{x_1, x_2, \ldots, x_N\}$, where N is the number of data points, and the goal is to find a partition of these points as $P = \{p_1, p_2, \ldots, p_N\}$ and then the center points of the partitions as $C = \{c_1, c_2, \ldots, c_k\}$. This will be done by optimizing an objective function [22]. Sum of squared errors (SSE) in “(4)” is among the most studied objective functions in the literature that we try to minimize during the clustering [21]–[23].

$$\mathrm{SSE} = \sum_{i=1}^{N} \|x_i - c_j\|^2 \tag{4}$$

Table III presents the average values of four measures across the four locations in the preprocessed dataset. The arrows in Table III explain how to interpret these values, indicating whether higher or lower values of each measure represent a stronger-tie network.

As illustrated in Table III, IS, RIS, and SS have higher values for stronger-tie networks, while OUT follows the opposite trend. To ensure consistency, we need to reverse the scale

TABLE II
THE THREE EXPLORED MEASURES USED WITH THE 2T FILTERING METHOD.

	Standard Deviation (STD)	**Interquartile Range (IQR)**	**Median Absolute Deviation (MAD)**
Functionality	Measures deviation from the mean	Measures spread of the middle 50%	Measures deviation from the median
Formula	$\sigma = \sqrt{\frac{1}{n}\sum(X_i - \mu)^2}$	$IQR = Q3 - Q1$	$MAD = b \times \text{med}(\lvert X - \text{med}(X)\rvert)$
Thresholds	Removes values ±3 × std from the mean	Removes values 1.5 × IQR lower/higher than Q1/Q3	Removes values 3 MADs from the median
Best for	Normally distributed data	Skewed or non-normal data	Heavy-tailed distributions
Robustness	No (sensitive to extreme values)	Yes (less affected by outliers)	Yes (most robust to extreme outliers)

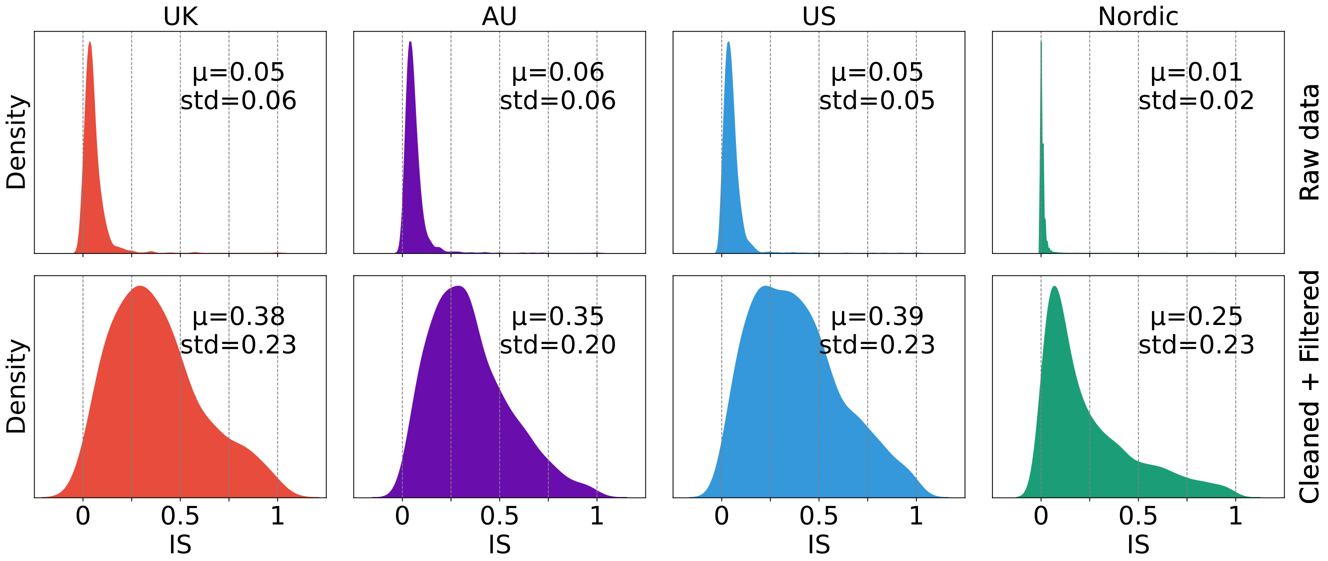

Fig. 4. Preprocessing impact on IS distributions. μ denotes mean, *std* stands for standard deviation, and data in each subplot is min-max normalized independently. Top: Distributions in the raw dataset across different locations. Bottom: Distributions after data cleaning and MAD-based filtering.

TABLE III
AVERAGE VALUES OF THE MEASURES UTILIZED FOR CLUSTERING. ARROWS REPRESENT EITHER THE HIGHER OR LOWER VALUES OF THE MEASURES REPRESENTING A STRONGER-TIE NETWORK.

	IS (↑)	**RIS (↑)**	**SS (↑)**	**OUT (↓)**
UK	2.30	3.10	0.01	27.08
AU	2.13	3.22	0.01	25.08
US	2.35	3.05	0.01	32.11
Nordic	1.61	1.89	0.02	35.80

of OUT so that higher values across all measures indicate stronger ties. We applied this reverse scaling separately for each location using "(5)" as follows:

$$\text{OUT}_{\text{new}} = \max(\text{OUT}) + \min(\text{OUT}) - \text{OUT} \qquad (5)$$

In "(5)", $max(OUT)$ and $min(OUT)$ indicate the maximum and minimum OUT values for each location, respectively.

Before clustering, we needed to normalize the data to ensure all measures are on the same scale. We used Z-score normalization because it does not allow bigger values to dominate the clustering process since they get transformed into standard deviations from the mean [24]. In addition, Z-score normalization is well-suited for distance-based algorithms such as k-means clustering [25]. In our analysis, we applied Z-score normalization globally to each measure, regardless of the country, to maintain consistency across the dataset. Since our study is unsupervised and clustering is performed on the complete dataset, it was appropriate to normalize globally rather than using statistics from a training subset.

After normalizing, we performed repeated k-means clustering using the Scikit-learn package in Python. We grouped the ego networks into four clusters: weak-tie, moderately-weak-tie, moderately-strong-tie, and strong-tie. The algorithm ran up to 1,000 iterations to refine cluster assignments and ensure convergence. To improve accuracy and avoid local optima, we initialized the centroids 30 times and selected the best result [26]. Table IV represents each cluster's centroids and the number of ego networks that fall within each cluster. To label clusters, we summed the values of each centroid's dimensions, sorted them in ascending order, and assigned the labels accordingly.

IV. RESULTS

Table V presents the clustering results based on the frequency of weak, moderately-weak, moderately-strong, and

TABLE IV
K-MEANS CLUSTERING RESULT WITH 30 RANDOM INITIALIZATIONS AND 1,000 ITERATIONS.

	Centroids				Counts	Label
	IS	RIS	SS	OUT		
Cluster 1	-0.53	-0.55	-0.54	-0.89	2,960	Weak
Cluster 2	1.43	-0.24	-0.35	-0.35	1,631	moderately-weak
Cluster 3	-0.47	-0.24	1.15	0.71	2,080	moderately-strong
Cluster 4	0.14	1.67	-0.15	0.50	1,502	Strong

strong tie labels across different locations. “Fig. 5” illustrates the proportion of each label within the four regions.

Nordic users maintain the lowest proportion of strong ties (11.9%) and the highest proportion of weak ties (39.3%), suggesting broad but overall shallow interactions within online communities for Nordic users. In contrast, Australian users demonstrate the strongest ties (36.0%) and the lowest ratio of weak ties (24%), pointing to deeper and more close-knit connections. The UK networks with nearly equal levels of strong (30.8%) and weak ties (31.1%) display a balance. US users lean toward weaker ties compared to the UK and AU, with 33.8% in the weak cluster and a lower percentage of strong ties (28.8%).

One might speculate that these differences in the strength of the US networks reflect cultural and communication norms, as Wellman et al. [27] noted that North American networks often support more dispersed, individual connections. Overall, these observations suggest that digital social structures vary by region and should be understood in specific cultural and social contexts.

TABLE V
CLUSTERING RESULTS ACROSS DIFFERENT LOCATIONS. EACH ROW INDICATES HOW MANY EGO NETWORKS FROM EACH LOCATION ARE LABELED AS WEAK, MODERATE, OR STRONG.

Location	Weak	moderately-weak	moderately-strong	Strong	Total
UK	119	117	29	118	383
AU	238	275	121	357	991
US	434	399	81	370	1,284
Nordic	2,169	840	1,849	657	5,515

A. Clustering validation

To evaluate the clustering quality and analyze the selection of the number of clusters (k), we employed three evaluation metrics: the Silhouette Score [28], the Calinski-Harabasz (CH) Score [29], and the WB-index [30]. “Fig. 6” presents these metrics plotted against varying values of k, ranging from 2 to 6.

The Silhouette Score ranges from -1 to 1, with values closer to 1 representing better cohesion and separation [28], peaks at $k = 4$, indicating well-separated and compact clusters at $k = 4$. Similarly, the CH Score, which evaluates cluster dispersion (higher values are better), is also highest at $k = 4$, suggesting that the data is best partitioned into four groups based on intra- and inter-cluster variance. WB-index (lower values better) reaches its minimum at $k = 5$, although the drop from $k = 4$ to $k = 5$ is slight.

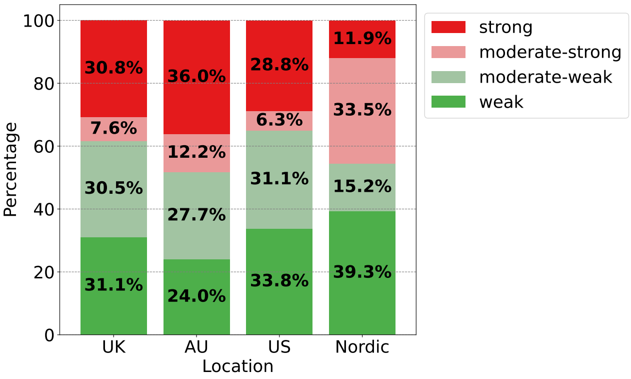

Fig. 5. Distribution of the clustering results by location.

The evidence across the three utilized metrics suggests the number of clusters as $k = 4$. In other words, having 4 clusters is expected to best capture structural distinctions within the data.

B. Gender vs. tie strength

In a study by Fränti et al. [31], the authors assigned gender labels to users in DSN corpora, excluding those from the Nordic region. The authors employed an ensemble method that considers name information, self-declaration of gender, and keywords in user profiles, and labeled accounts into four gender categories: male (M), female (F), non-binary (NB), and uncategorized (U).

In this paper, our focus is on the networks as data units rather than the users themselves. Therefore, we consider the network’s gender label based on the gender of its central node (the ego) labeled by Fränti et al. [31]. For instance, when we refer to a ”male network,” it denotes a network whose central user was labeled as male. Among the gendered networks, 2,550 networks passed the preprocessing step (see Section 2.2) and were included in this study. Table VI illustrates the distribution of these networks across different locations.

Focusing only on male and female networks across different locations, “Fig. 7” displays that the share of male networks in the UK, AU, and US is consistently larger than female networks across all four tie strength clusters. This finding aligns with the previous research [32]–[34], which reported

TABLE VI
DISTRIBUTION OF GENDERED USERS IN DSN CORPORA IN DIFFERENT LOCATIONS.

	Male	Female	Uncategorized	Non-binary	Total
UK	199	95	72	0	366
AU	494	220	227	3	944
US	621	342	273	4	1,240
Total	1,314	657	572	7	2,550

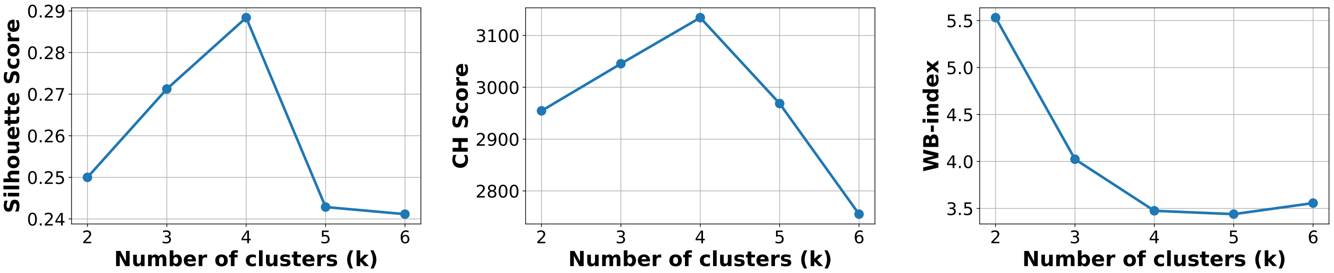

Fig. 6. Clustering validation metrics across different number of clusters (k).

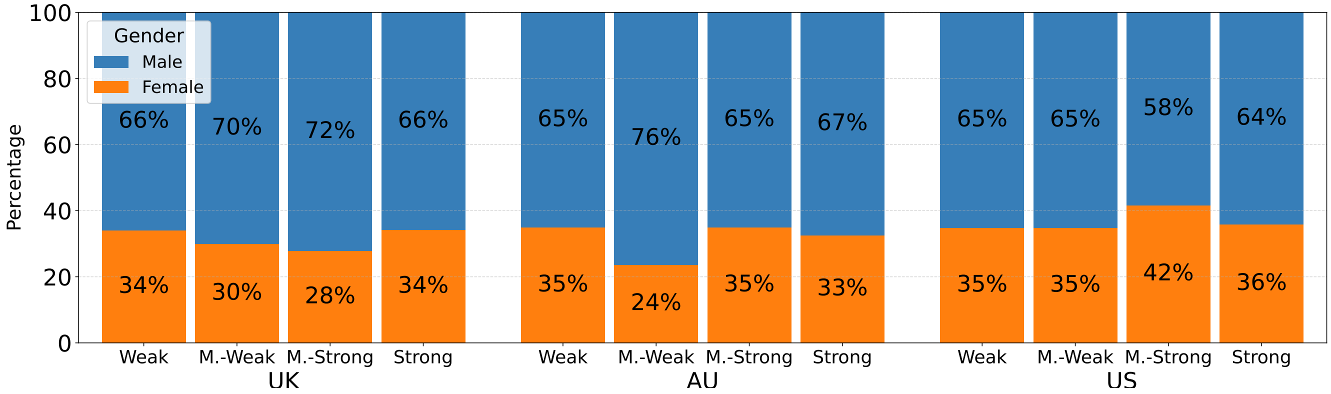

Fig. 7. The percentage distribution of male and female networks in different tie strength clusters and locations.

that Twitter users are mostly male compared to the general population.

V. Conclusion

In this study, we developed a clustering-based approach to classify digital ego networks from Twitter into four strength categories: weak, moderately-weak, moderately-strong, and strong. We calculated four interactional and structural measures studied by Fatemi et al. [13] from each network in the dataset and transformed them into a four-dimensional space for clustering.

The method we proposed addressed key limitations in previous research by implementing an automated and scalable pipeline that relies on four quantitative measures: interaction strength (IS), relative interaction strength (RIS), social similarity (SS), and outlier (OUT). Although the study uses data from one application (old Twitter), the method is applicable to any social media platform where user networks can be constructed.

Our results revealed consistent regional differences in network structures, with Nordic users indicating a preference for weak-ties. In contrast, users from Australia, the UK, and the US were more likely to form strong-tie networks, with Australia demonstrating the highest share of strong-tie networks and the fewest weak-tie networks. This may be influenced by factors such as how data is collected or exchanged.

In addition, our results demonstrate that compared to female networks, male networks consistently held the majority of networks across all tie categories and all locations. Regional patterns also highlighted cultural variations in digital networking behaviors, particularly in the Nordic region, where individuals tend to have broader but looser social connections.

One key limitation in this study is the lack of temporal dynamics. In other words, the dataset does not accommodate time-dependent analyses, and our analyses are static, not accounting for how tie strength might change over time. As part of our future work, we are constructing a comparable network dataset from another online social network that incorporates temporal information, enabling the examination of changes in tie strength over time.

Our proposed approach enables the development of a scalable method for detecting tie strength in online social networks. Our findings can be applied to improve future studies that model information diffusion in fundamental research, while outside academia, the method could be used to improve content personalization and possibly targeted advertising. Additionally, our approach supports sociocultural research by revealing regional and gender-based patterns in online social interactions.

Data and code availability

The code supporting this study is publicly available on GitHub at: https://github.com/uef-machine-learning/clustering-networks-by-tie-strength

To preserve privacy, the underlying datasets are not published; however, they can be requested for academic research purposes by contacting the corresponding author.

Acknowledgment

This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Research Council of Finland under grant numbers 345640, 358725, 367757 (FIRI 2022–29) and 364048 (COMET Academy project for 2024–2028). We also would like to thank CSC – IT Center for Science for providing access to their supercomputers, which were essential for data storage and code execution required for this project. The project also received early-stage funding from the Center for Data Intensive Sciences and Applications (DISA) at Linnaeus University.

References

[1] M. S. Granovetter, "The strength of weak ties," American J. Sociology, vol. 78, no. 6, pp. 1360–1380, 1973.

[2] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 519–528, doi: 10.1145/2187836.2187907.

[3] P. S. Park, J. E. Blumenstock, and M. W. Macy, "The strength of long-range ties in population-scale social networks," Science, vol. 362, no. 6421, pp. 1410–1413, 2018, doi: 10.1126/science.aau9735.

[4] M. Del Tredici and R. Fernández, "The road to success: Assessing the fate of linguistic innovations in online communities," in Proc. 27th Int. Conf. Comput. Linguistics, E. M. Bender, L. Derczynski, and P. Isabelle, Eds. Santa Fe, NM, USA: Assoc. Comput. Linguistics, Aug. 2018, pp. 1591–1603. [Online]. Available: https://aclanthology.org/C18-1135/

[5] M. Laitinen, M. Fatemi, and J. Lundberg, "Size matters: Digital social networks and language change," Frontiers Artif. Intell., vol. 3, May 2020, Art. no. 46, doi: 10.3389/frai.2020.00046.

[6] D. Centola and M. Macy, "Complex contagions and the weakness of long ties," American J. Sociology, vol. 113, no. 3, pp. 702–734, Nov. 2007, doi: 10.1086/521848.

[7] G. Walker, B. Kogut, and W. Shan, "Social capital, structural holes and the formation of an industry network," Organization Sci., vol. 8, no. 2, pp. 109–125, Mar./Apr. 1997, doi: 10.1287/orsc.8.2.109.

[8] K. B. Wright and C. H. Miller, "A measure of weak-tie/strong-tie support network preference," Commun. Monogr., vol. 77, no. 4, pp. 500–517, Dec. 2010, doi: 10.1080/03637751.2010.502538.

[9] H. R. Bernard, T. Hallett, A. Iovita, E. C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M. J. Salganik, T. Saliuk, O. Scutelniciuc, G. A. Shelley, P. Sirinirund, S. Weir, and D. F. Stroup, "Counting hard-to-count populations: The network scale-up method for public health," Sex. Transm. Infect., vol. 86, suppl. 2, pp. ii11–ii15, Nov. 2010, doi: 10.1136/sti.2010.044446.

[10] Statista, "Number of worldwide social network users (Statista, No. 278414)," *Statista*, n.d. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. [Accessed: Aug. 14, 2025].

[11] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," Proc. Nat. Acad. Sci. USA, vol. 104, no. 18, pp. 7332–7336, May 2007, doi: 10.1073/pnas.0610245104.

[12] R. Goel, S. Soni, N. Goyal, J. Paparrizos, H. Wallach, F. Diaz, and J. Eisenstein, "The social dynamics of language change in online networks," in Social Informatics, E. Spiro and Y.-Y. Ahn, Eds. Cham, Switzerland: Springer Int. Publishing, 2016, vol. 10046, pp. 41–57, doi: 10.1007/978-3-319-47880-7_3.

[13] M. Fatemi, M. Laitinen, and P. Franti, "Computer-mediated communication and networks: Quantifying tie strength," In Special Issue on Computer-Mediated Communication Corpora, Language@Internet, 2026 (Forthcoming), doi: 10.xxxxx.

[14] J. Scott and P. J. Carrington, *The SAGE Handbook of Social Network Analysis*. London, UK: SAGE Publications, 2014. doi: 10.4135/9781446294413

[15] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," Social Netw., vol. 25, no. 3, pp. 211–230, Jul. 2003, doi: 10.1016/S0378-8733(03)00009-1.

[16] D. M. Feehan, V. Hai Son, and A. Abdul-Quader, "Survey methods for estimating the size of weak-tie personal networks," Sociol. Methodol., vol. 52, no. 2, pp. 193–219, Aug. 2022, doi: 10.1177/00811750221109568.

[17] K. Rajkumar, G. Saint-Jacques, I. Bojinov, E. Brynjolfsson, and S. Aral, "A causal test of the strength of weak ties," Science, vol. 377, no. 6612, pp. 1304–1310, Sep. 2022, doi: 10.1126/science.abl4476.

[18] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," WIREs Data Mining Knowl. Discovery, vol. 1, no. 1, pp. 73–79, Jan./Feb. 2011, doi: 10.1002/widm.2.

[19] J. Yang, S. Rahardja, and P. Fränti, "Outlier detection: How to threshold outlier scores?," in Proc. Int. Conf. Artif. Intell., Inf. Process. Cloud Comput., 2019, pp. 1–6, doi: 10.1145/3371425.3371427.

[20] R. Wilcox, Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction, 1st ed. Boca Raton, FL, USA: CRC Press, 2011.

[21] M. Rezaei and P. Fränti, "Can the number of clusters be determined by external indices?," IEEE Access, vol. 8, pp. 89239–89257, 2020, doi: 10.1109/ACCESS.2020.2993295.

[22] L. Rokach and O. Maimon, "Clustering methods," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. Boston, MA, USA: Springer US, 2005, pp. 321–352, doi: 10.1007/0-387-25465-X_15.

[23] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, "Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the elbow method," J. Phys.: Conf. Ser., vol. 1361, no. 1, Art. no. 012015, Nov. 2019, doi: 10.1088/1742-6596/1361/1/012015.

[24] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[25] D. Steinley, "K-means clustering: A half-century synthesis," Br. J. Math. Stat. Psychol., vol. 59, no. 1, pp. 1–34, May 2006, doi: 10.1348/000711005X48266.

[26] P. Fränti and S. Sieranoja, "How much can K-means be improved by using better initialization and repeats?," Pattern Recognit., vol. 93, pp. 95–112, Sep. 2019, doi: 10.1016/j.patcog.2019.04.014.

[27] B. Wellman, A. Quan-Haase, J. Boase, W. Chen, K. Hampton, I. Díaz, and K. Miyata, "The social affordances of the Internet for networked individualism," J. Comput.-Mediated Commun., vol. 8, no. 3, Art. no. JCMC834, Apr. 2003, doi: 10.1111/j.1083-6101.2003.tb00216.x.

[28] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[29] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," Commun. Stat., vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

[30] Q. Zhao and P. Fränti, "WB-index: A sum-of-squares based index for cluster validity," Data Knowl. Eng., vol. 92, pp. 77–89, Oct. 2014, doi: 10.1016/j.datak.2014.07.008.

[31] P. Fränti, J. Järviö, M. Salimi, I. Taipale, M. Laitinen, R. Albicker, C. Nie, M. Fatemi, and P. Rautionaho, "Beyond names: How to label gender automatically in CMC data?," in Proc. 12th Int. Conf. CMC and Social Media Corpora for the Humanities (CMC-Corpora), Bayreuth, Germany, 2025.

[32] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist, "Understanding the demographics of Twitter users," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 5, no. 1, Art. no. 1, Jul. 2011, doi: 10.1609/icwsm.v5i1.14168.

[33] T. Laor, "My social network: Group differences in frequency of use, active use, and interactive use on Facebook, Instagram and Twitter," Technol. Soc., vol. 68, Art. no. 101922, May 2022, doi: 10.1016/j.techsoc.2022.101922.

[34] M. Macedo and A. Saxena, "Gender differences in online communication: A case study of Soccer," arXiv preprint arXiv:2403.11051, Mar. 2024, doi: 10.48550/arXiv.2403.11051.