SN Computer Science (2025) 6:815
https://doi.org/10.1007/s42979-025-04353-y

ORIGINAL RESEARCH

Detecting Connectivity Patterns in Nordic Twittersphere by Cluster Analysis

Masoud Fatemi[1,3] · **Sami Sieranoja**[1] · **Mikko Laitinen**[2,3] · **Pasi Fränti**[1]

Received: 10 October 2024 / Accepted: 29 August 2025

Abstract
We analyze Nordic social media users by clustering them based on their connections on Twitter. The data consists of 15,794 users in the five Nordic countries: Finland, Sweden, Norway, Denmark, and Iceland. We first create an undirected graph from the friendship relations (mutually following each other), then divide the graph into five clusters using a recent M-algorithm, and finally compare the results to users' locations. The results demonstrate that the users are strongly clustered according to their home country. There is surprisingly little interaction across the countries despite the fact that they are, except for Iceland, physically close to each other and have cultural and linguistic similarities. The main language of the four countries belongs to the Germanic languages, while Finnish is typologically distinct. We further explore content from users in each country, analyzing its alignment with connectivity patterns. Our findings reveal a discrepancy between user-generated content similarity in the Nordic region and their connectivity patterns.

Keywords Clustering · Twitter users · Social networks · Nordic countries · Community detection · Graph clustering

Introduction

This study focuses on clustering social media users' locations based on their connections. The study uses data from X (previously known as Twitter). As is widely known, it is a prominent social media platform that has a significant influence on our daily lives [1]. This is primarily due to the diverse information (news, opinions, and personal experiences) that users share with others on the application [1]. Through interactional links (mutually following each other), users create social networks that can be helpful in analyzing societal behaviors [2], information dissemination [3], and impact on other users and public debates [3]. Within these networks, virtual communities can also be identified. The abundance of data on Twitter has meant that it has become a vast repository that is utilized in both applied fields, such as marketing [1], and in extracting novel insights and knowledge in fundamental research [2, 4].

For over a decade, researchers have studied national or language-specific Twitter communities, analyzing aspects like network structures, clustering, and user interaction dynamics [5–7]. In [5], authors studied a follower/followee network of 120,000 accounts in Australia (called *Australian Twittersphere*). They used Australian-themed hashtags verified by the time-zone settings of the users (unique to Australia). Their study delivers insights into the spread of hashtags on Twitter and highlights the discovery of a significant portion of Australian Twitter users, paving the way for innovative data collection methods.

The authors in [7] compared a single-day activity to a long-term activity using 177,000 unique accounts in the Australian Twittersphere. They observed more diversity in

✉ Masoud Fatemi
masoud.fatemi@uef.fi

Sami Sieranoja
sami.sieranoja@uef.fi

Mikko Laitinen
mikko.laitinen@uef.fi

Pasi Fränti
Pasi.franti@uef.fi

[1] Machine Learning Group, School of Computing, University of Eastern Finland, Joensuu, Finland

[2] School of Humanities, University of Eastern Finland, Joensuu, Finland

[3] Center for Data Intensive Sciences and Applications, Linnaeus University, Växjö, Sweden

Published online: 10 September 2025

the single-day activity patterns. They also highlighted the limitations of hashtag-driven methods. Van Geenen et al. in [8] made an attempt to analyze one week of activity on Twitter by detecting accounts of politicians, media organizations, and journalists. However, no significant findings were reported.

Kwak et al. in [9] investigated Twitter's dynamics by examining a network among 41.7 million users. They used PageRank and the number of followers to identify influential users. The results revealed unique characteristics, such as non-standard follower distributions and fast information diffusion, primarily through retweets.

In a series of papers [10–12], Münch et al. analyzed the German and Italian Twitter networks. The first paper introduces a sub-sampling method based on rank-degree [10]. The authors focus only on nodes with higher connection degrees. In the follow-up paper, they examined the relationship between the Italian and German Twitter communities using a sample of 14,685 nodes extracted based on the language of the Tweets [11]. Their third paper showed that the sub-sampling approach was able to identify the top influential accounts in the German Twittersphere based on 1–10% sub-sample size [12].

The main limitation of these studies is that analyzing large networks requires good tools. Sub-sampling is one possibility to reduce the size of the data that would be impractical to analyze manually. However, clustering would be more appropriate in summarizing extensive amounts of data [13]. For example, multimorbidity graph was constructed from 58 million patient diagnoses in Finland and then partitioned into diagnosis clusters [14]. The summarization by the clustering made it easier to analyze the content, which would have been an overwhelming task if examined the full data as such. Clustering has been in various fields, such as health science, online marketing, and transportation [14–16].

In social networks, clustering has been employed to detect communities [13, 17–19]. The authors in [20] applied the Louvain clustering algorithm for the Australian Twittersphere to extract 30 major clusters. They compared the thematic content of the clusters and found a shift from a technology-centric base to more diverse ones encompassing sports, politics, and celebrity culture. The same clustering algorithm was applied to the Norwegian Twittersphere selected based on the interface language and the profile location information in [6]. The study focused on the echo chamber phenomenon, but the extracted clusters revealed very little evidence for it.

In [21] authors presented a hierarchical clustering algorithm with an information-theoretic clustering criterion focusing on the hierarchical aspect of the network. Peixoto introduced another information-theoretic clustering method based on the minimum description length principle to estimate the number of clusters [22]. If the data had natural clusters, this approach could potentially find their correct number. A follow-up paper by Peixoto provides a theoretical background of clustering via extensive discussion of several myths that sometimes appear in the literature [23]. We fully agree with the arguments made in the paper.

In this paper, we apply cluster analysis to analyze the Nordic Twittersphere. Similar to [20], we use the mutual follower/followee relationship with the assumption that a mutual relationship creates a stronger link than a simple following relation. We use a recent M-algorithm with conductance criterion, which has been shown to be more robust on Lancichinetti data than the widely used Louvain algorithm [19]. We do not claim the M-algorithm as a novel contribution; however, to our knowledge, this is the first time it has been applied to community detection in Twittersphere data.

Contrary to the Australian Twittersphere [5], we do not have an obvious shortcut like the unique time zone to select the Nordic users. Instead, we rely on the geo-tagged data collected in the Nordic Tweet Stream (NTS) project [24]. Although location data is not always available due to privacy or other concerns [25], some Twitter users still share their locations to provide us with a large dataset on which to base our analysis. Without such geo-tagged database, researchers would need to develop methods to infer location.

We compare the clustering result with the physical locations of the users. Specifically, we aim to determine whether a correlation exists between the clustering results and the users' home countries in the Nordic region. The focus on Nordic countries is justified given that the five countries share substantial socio-cultural similarities, and the region has been suggested to be a "laboratory" for research into the contexts and consequences of globalization and mobility [26]–[27].

The process we followed involve several steps. We first collected tweets from the Nordic region between November 2016 and November 2022. We selected users whose tweet locations matched their profile locations in the five Nordic countries. We then excluded users who had location-tagged tweets in another country than their home country indicated by their profile. The resulting graph was then clustered using a state-of-the-art graph clustering [19]. We opted for five clusters representing the five countries in the data, but we also examined the impact of adding a sixth cluster.

The study seeks to answer the following research questions. First, we aim to determine how accurately the clustering results align with the country division of the users. Second, we explore whether there are any hidden or undiscovered clusters that were not initially apparent. Third, we investigate which clustering criterion is the most appropriate for the given data. We make a brief content analysis of the country clusters on their use of hashtags although we are

aware of its limitations. Our primary goal is to explore the existence of clusters, not extensive content analysis.

While previous studies have explored Twitter networks in national contexts, our work contributes a novel regional perspective by analyzing the Nordic Twittersphere as a unified yet culturally diverse space. The clear correspondence between social connectivity and national borders in our results, despite geographic proximity and shared cultural features, reveals new insights into how digital communities mirror offline identities. This approach not only deepens understanding of social clustering in a multilingual, multi-country context, but also provides a replicable framework for analyzing regional networks elsewhere.

While authors in [5] have mapped national Twitter networks and observed that users often cluster based on geographic and cultural lines, our study extends this analysis to the Nordic Twittersphere using more recent data from 2022. By applying a state-of-the-art clustering algorithm to a large-scale [19], multilingual dataset, we aim to uncover the nuances of regional digital communities in the Nordic context, offering comparative insights across multiple countries with shared and divergent cultural traits.

The structure of the paper is as follows. Section "Nordic Twitter Data" documents the data collection process and summarizes the properties of the data. Section "Clustering" reviews previous research on network clustering and their result analysis. It also details the selected clustering algorithm and studies the effect of the different objective functions. The clustering results are discussed in Sect. "Results". Sect "Conclusions" concludes the paper and suggests potential future research.

Nordic Twitter Data

Our main data source is the Nordic Tweet Stream (NTS), which has been collected continuously since November 2016 [24]. NTS is a constantly growing dataset that includes geolocation-enabled tweets from the Nordic countries from November 2016 up to the present [24]. For this study, we selected users who had tweets between November 1, 2016, and November 31, 2022, resulting in a dataset called Nordic Twitter Network (hereafter *NTN-2022*). This dataset comprised a total of 691,521 user accounts.

The time frame was selected to cover only the era before the change of Twitter to X. We wanted to minimize the impact of external factors such as ownership changes can have on a social media platform and its user communities. Our choice also allows the data to be used later for comparative studies between Twitter and X.

We opted to use geo-location as the selection criterion for the users in the Nordic Twittersphere for two reasons. First, in this way, we directly address the challenge of targeting research findings to specific geographical areas. This is particularly valuable for understanding regional nuances and how local contexts influence Twitter interactions. Second, a country hashtag such as #Finland gives no guarantee that the person is from Finland. We do not have similar unique time-zone verification as Australia. Limitations of Hashtag as a selection criterion have been widely noted in literature [28–30].

Instead, we rely on the geo-tagged data collected in the Nordic Tweet stream (NTS) project [24]. This may filter out many users and have a strong sub-sampling effect on the data with possible bias. However, the selected users are likely more knowledgeable than those not using geo-location. This aligns well with the other researchers focusing on expert users [31].

Location Information

Twitter offers two types of locations. The first is the user's self-reported location (text field) in the Bio section. This field is not standardized and may be inaccurate, as users can enter any location they choose, even a fictional one [32]. The second is the geolocation feature, which can be added to every tweet by users who enable this option in their profile settings. This location is provided automatically, and it is in a standardized format, including the latitude, longitude, and the country code. NTS consists of tweets that have this secondary location in one of the five Nordic countries: Finland, Sweden, Norway, Denmark, and Iceland.

Nordic Twitter Network

The process of creating the *NTN-2022* dataset included the following steps: (1) user extraction, (2) user labeling, (3) user filtering, and (4) collecting the tweets in the entire network, see Fig. 1. In the first step, we extracted all users who had tweets included in the NTS between November 2016 and November 2022. In the second step, we labeled all users based on the country they tweeted from. Users form five distinct sets representing the five Nordic countries.

Users who had tweets from more than one country were excluded. Including such users may introduce inconsistencies, making it difficult to accurately categorize and analyze their generated content. We intend to focus exclusively on topics or discussions in Nordic countries. Including users who had tweeted from more than one Nordic country might dilute the data set intended geographical focus. Moreover, excluding users with tweets spanning multiple countries improves the data quality and simplifies data analysis and interpretation for more straightforward comparisons and insights. In this step, 88,381 accounts were excluded.

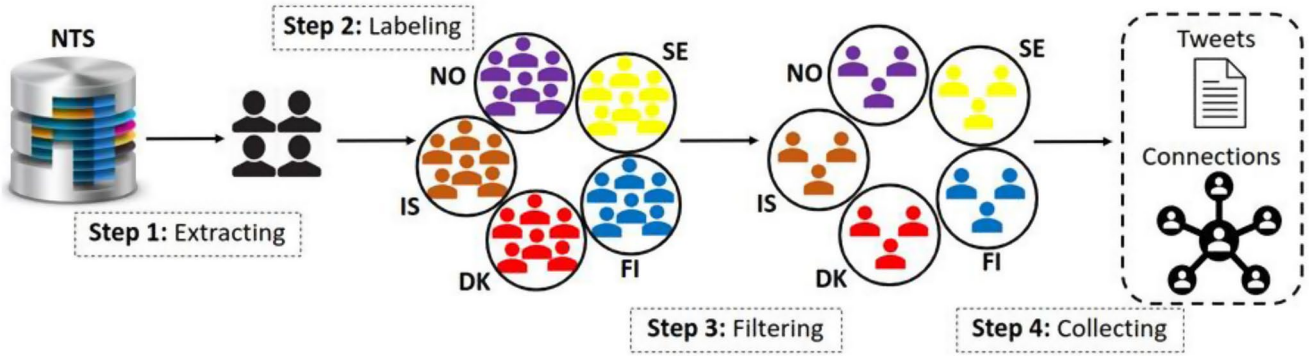

Fig. 1 An overview of NTN construction steps

Table 1 Statistics of the *NTN-2022* dataset

Country	Nodes	Edges	Average Degree	Density (·1000)	Tweets
Finland	5,872	27,855	4.8	0.83	2,392,135
Sweden	6,871	18,555	2.7	0.39	6,137,063
Denmark	1,490	3,434	2.3	1.55	1,275,974
Norway	1,390	2,357	1.7	1.22	1,613,866
Iceland	261	561	2.1	8.27	178,925
NTN	15,794	54,027	3.4	0.22	11,597,963

In the third step, we filtered out abnormal users and users with some uncertainty in their location. Specifically, we only selected users who self-reported their location in their profile, and it matched to the location of their tweets. Consequently, another 484,064 accounts were excluded from the data set. In addition, we excluded verified accounts that (at the time of data collection) were common for celebrities and politicians so that the network consists of mainly real genuine people.

As for the network size, we assume the size of a typical human network to be over 30–50 [33], and 150–200 is the estimated average human network size that one can maintain and interact with [34]–[35]. To adjust to these assumptions of human networks, we filtered out users who had more than 500 contacts. We also excluded the top 1% (very active) and bottom 1% (very passive) accounts based on the number of tweets. As a result, the initial dataset of 691,521 users was reduced to 37,057 users.

Once the user list was finalized, we collected the connections between these users as well as their tweets (up to 3,200 latest messages, excluding retweets). Directed links were established between the users based on the interactional relationships. Isolated nodes and smaller disjoint sub-networks parts were excluded, and only the largest connected component was kept as the final *NTN-2022* dataset.

Table 1 summarizes the statistics of the *NTN-2022* dataset. The network includes 15,794 nodes and 54,027 links. Most users are from Finland (37%) and Sweden (43%). Iceland is the smallest country in this data (2%). Density is the number of edges relative to all possible edges in a complete graph. For readability purposes, density values are multiplied by 1,000. Figure 2 illustrates a sample graph from *NTN-2022* with 3,273 nodes and 13,483 edges drawn by the Gephi open-source network analysis software [36] using the Force Atlas 2 algorithm [37].

Clustering

We next describe the clustering algorithm and the components it includes and explain the choices behind each of them.

In the standard clustering problem, we have a set of points as $X = \{x_1, x_2, \ldots, x_N\}$, and the goal is to find the 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 happens by minimizing an objective function such as the sum of squared errors in (1) [38]:

$$SSE = \sum_{i=1}^{N} \| x_i - c_j \|^2 \quad (1)$$

In graph clustering and community detection, the input data is a graph consisting of a set of nodes and edges. The goal is to identify subsets of nodes (called clusters or communities) so that in each subset, nodes are strongly connected within the set and loosely connected to nodes outside the set [39]. As in the standard clustering problem, an objective function needs to be optimized.

Existing Approaches

Graph clustering algorithms can be categorized into three approaches: *agglomerative*, *divisive*, and *iterative* [18]. Agglomerative algorithms merge nodes recursively until the desired number of clusters is reached [40]. Divisive algorithms do the opposite and remove connections until a desired number of isolated components are reached

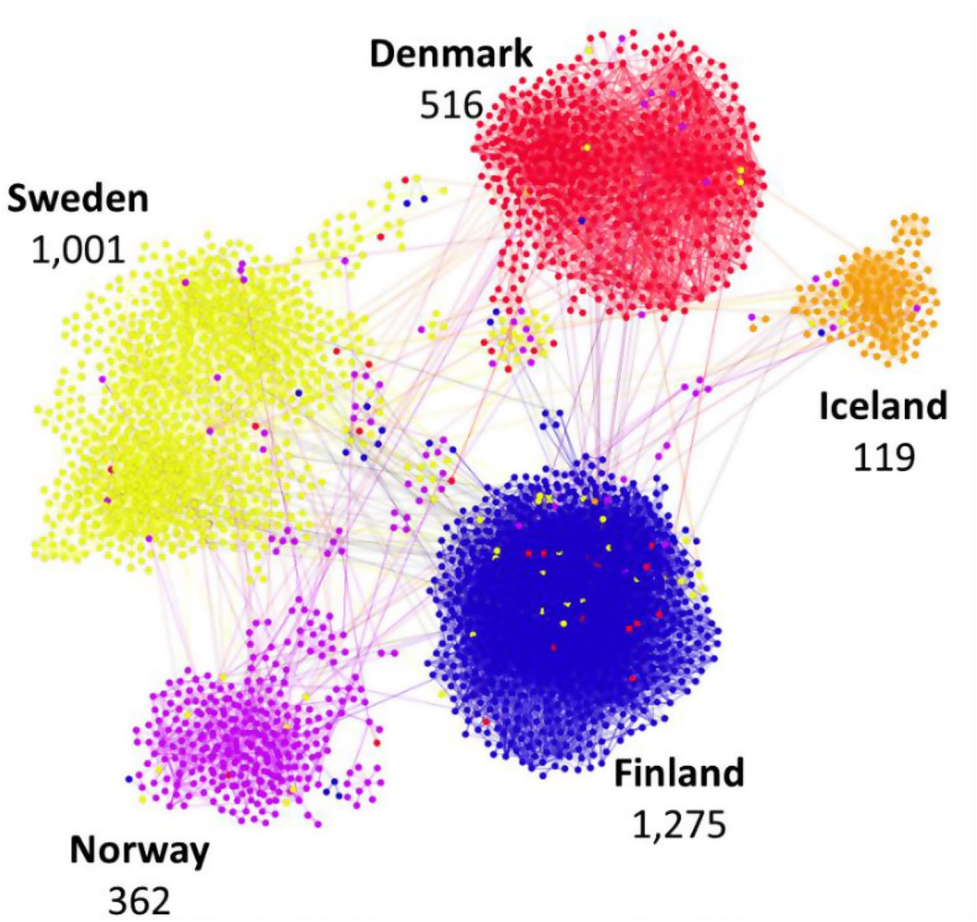

Fig. 2 A sample of *NTN-2022* dataset visualized according to the country of the users

[41]–[42]. Iterative algorithms use an objective function which is either minimized or maximized via local optimization steps [42]–[43].

Users may also belong to multiple communities as they naturally overlap. The seed expansion method in Whang et al. grows communities in an overlapping manner [44]. In [44] the authors utilized the conductance cost function and tried to optimize it.

In graph theory, a highly irregular graph is a graph in which for each node, all the neighbors (nodes directly connected to it) have different degrees [45]. Karypis and Kumar in [46] aimed to partition irregular graphs using a three-step process: collapsing nodes and edges (coarsen), detecting communities in the coarsen graph through a seed expansion, and refinement of the coarsen graph. The algorithm forms balanced clusters, which is not the case with *NTN-2022* data (see Fig. 2).

The authors in [47] developed a machine learning model for simultaneous graph embedding and clustering. Graph embedding refers to transforming complex and nonlinear nodes and edges into a low dimensional Euclidean space (usually vector space) while preserving the main criteria of the graph. In social network embedding, preserving community membership is a priority. The model uses a parameter to control the proximity of nodes during the transformation.

Louvain algorithm is the most common algorithm for detecting communities in social networks, possibly because it has been implemented in Gephi [18]. It is an agglomerative algorithm that optimizes modularity as the cost function. *Modularity* is a measure that compares the number of edges within a cluster to the expected number if the edge weights of the same nodes were randomly distributed (null hypothesis) [48]. Clustering is considered good if there are more edges within the clusters than expected. The algorithm is fast.

Embedding-based methods, such as DeepWalk [49] and GraphSAGE [50], offer an alternative approach by learning node representations that can then be clustered using standard algorithms like k-means. However, our method directly clusters the graph based on link structure, which is more suitable for our analysis.

The authors in [51] argue that in real-world graphs, there might be several partitions that are close to the global optimum. They discussed that an expert could select the best among the several good partitions using their domain knowledge. However, in the case of large communities, it would be an overwhelming task to do. Hence, they proposed to split the large partitions into smaller and similar parts to provide an abstract interpretation and adequate information about the primary partition.

It is possible to obtain more information from a single network by repeating measurements over different time periods [52]–[53]. A stochastic framework and a Gibbs sampling procedure have been used in [54] to cluster similar structures within a population of networks instead of focusing on a single network.

We will use a newly proposed graph clustering algorithm due to its good clustering accuracy. It was shown to provide significantly more accurate results than the other algorithms tested in [19] including the widely used Louvain algorithm. It is important to have an accurate and reliable clustering algorithm so that we can focus on the clustering results instead of needing to worry about algorithm performance or artifacts.

M-Algorithm

The algorithm is called M-algorithm (see Algorithms 1–2) [19]. It is a direct derivative of the k-means algorithm adapted for graphs with an additional split-and-merge step. The algorithm has several advantages [55]. First, it is computationally efficient and relatively simple compared to many other algorithms used for graph clustering. Second, it is versatile in the sense that it can be applied for several types of objective functions, depending on the application. For example, it can be utilized to detect clusters with either balanced or unbalanced cluster sizes.

K-means finds the best cluster (one that minimizes objective function) indirectly by calculating distances to the mean vectors of the clusters and selecting the nearest according to Euclidean distance. This makes it fast but works only for the SSE objective function. It is not possible to apply k-means directly to networks or graphs without embedding the data into vector space. This would degrade the clustering quality by adding an extra approximation layer to the process.

The M-algorithm finds the nearest cluster for a given node by estimating the change (delta) on the objective function when switching the node from one cluster to another [19]. This can be done fast because the delta depends only on the neighbors of the node. This delta approach makes it possible to use the M-algorithm with many other objective functions.

The algorithm includes two main phases [19]. The first phase, called K-algorithm (Algorithm 2), works in a similar way as k-means. It first constructs an initial clustering by growing clusters in random locations. The partitions are then gradually fine-tuned by switching nodes to another partition if there exists one that improves the cost function. The algorithm repeats these phases as long as there are any changes in the clusters.

The K-algorithm always converges to a local optimum, which is sometimes far from the global optimum. To improve the result, the second phase is implemented. It follows a merge-and-split strategy. First, a random pair of clusters is merged. Then, a random cluster is split. Finally, the new clustering is fine-tuned using the K-algorithm. The new clustering is kept if it improves the objective function over the current best candidate. This process is repeated a user specified number of times, allowing a flexible compromise between clustering quality and processing time.

```
Algorithm 1: M_algorithm(graph,k,R)
INPUT:
graph (with N nodes)
k = number of clusters
R = number of repeats
 1 bestClustering = K_algorithm(graph,NULL,k)
 2 FOR i=1:R
 3   newClu = bestClustering
 4
 5   // Merge two random clusters
 6   (A,B) = Choose a pair of random clusters
 7   newClu = merge(newClu, graph, A,B)
 8
 9   // Split one random cluster
10   cluToSplit = Choose one random cluster
11   newClu = split(newClu, graph, cluToSplit)
12
13   // Finetune using K-algorithm
14   newClu = K_algorithm(graph,newClu,k)
15   IF cost(graph,newClu) > cost(graph,cluster) // improvement
16      bestClustering = newClu
17 RETURN bestClustering
```

```
Algorithm 2: K_algorithm(graph,k,cluster)
INPUT:
graph (with N nodes)
k = number of clusters
cluster = initial clustering (optional)
 1 IF cluster == NULL
 2   cluster = InitialPartition(graph,k)
 3 DO
 4   changed = 0
 5   FOR i=SHUFFLE(1:N) // Process nodes 1..N in random order
 6     old = cluster[i]
 7    newpart = 1
 8    bestdelta = INF
 9    FOR j=1:k // Loop all clusters
10      delta = changeInCost(i,cluster[i],j)
11      // If moving node i to cluster j improves cost
12      IF delta < bestdelta
13        bestdelta = delta
14        newpart = j
15    IF new != old
16      changed += 1
17      cluster[i] = new
18 WHILE changed > 0
19 RETURN cluster
```

Objective Functions

We consider three objective functions: conductance (CND) [56], inverse internal weight (IIW) and mean internal weight (MIW) introduced in [19]. They are defined as follows:

$$CND = \frac{1}{k}\sum_{i=1}^{k}\frac{E_i}{T_i} \tag{2}$$

$$IIW = \frac{M}{k^2}\sum_{i=1}^{k}\frac{1}{W_i} \tag{3}$$

$$MIW = \frac{1}{k}\sum_{i=1}^{k}\frac{W_i}{n_i} \tag{4}$$

where n_i is the size i^{th} cluster, k is the number of clusters, W_i is the sum of internal weights in cluster i, E_i is the sum of external weights from cluster i, and T_i is the total weight of the edges connecting to the nodes in cluster i $(E_i + W_i)$. All the objective functions consist of individual components for each of the k clusters. The total objective function value is calculated as the average of these.

Based on Formula 2, for each state, the CND value (between 0 and 1) equals the summation over all the weights of all external edges from each cluster divided by the total weight of the nodes in that cluster. A small value for conductance represents a good clustering. Minimizing conductance denotes a lower value for the sum of the external weights (E_i) and a higher value for the sum of internal weights (W_i). Conductance also avoids creating overly small clusters.

The IIW objective function, see Formula 3, has a value in the $[1, \infty]$ range and is the summation of internal weights for each cluster (W_i) multiplied by a constant. Like CND, minimizing IIW leads to better clustering results. For example, in the case of optimal clustering where k completely separated and balanced clusters are calculated from the network, all W_i would equal M/k, and the IIW value would be 1.

The MIW proposed in [19] is the weighted version of the objective function introduced in [57]. Based on Formula 4, it normalizes the internal weights for each cluster (W_i) by dividing by the cluster size (n_i). The MIW objective function must be maximized to result in more disjoint clusters. Maximizing MIW tends to form small dense clusters and one large "garbage cluster" for non-dense parts of the graph.

Results

In this section, we take an in-depth look at the *NTN-2022*. We first examine the actual network and the links between and within countries. We then apply the Malgorithm discussed in Sect. "Clustering" and evaluate the impact of the objective function on the clustering results. Lastly, we consider data created by users, namely hashtags, to determine the similarity of content produced by users from various

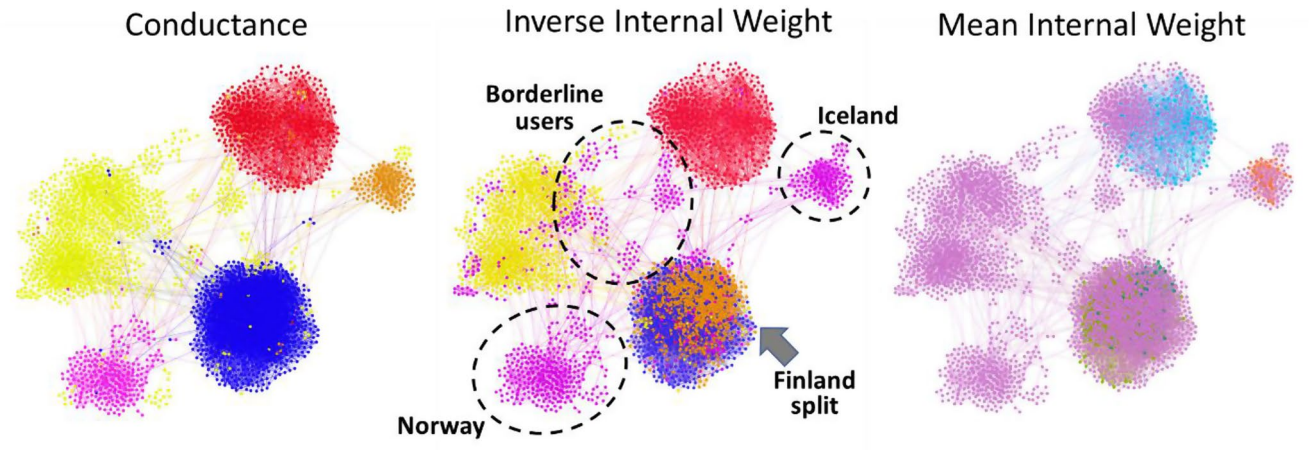

Fig. 3 Clustering with three different objective functions

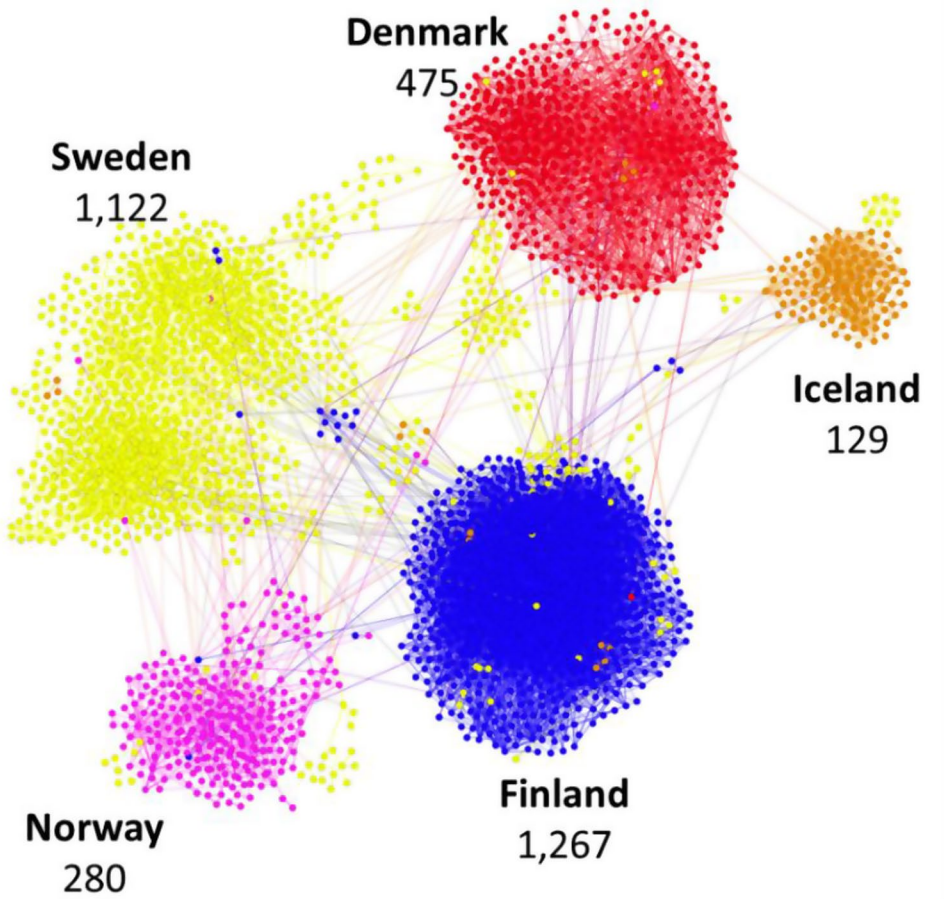

Fig. 4 Detected five clusters using the M-algorithm with conductance function. The result matches very closely the home countries of the users

countries and explore the similarity in content and connection patterns between countries.

Clustering Objective

The visualization in Fig. 2 suggests that the data is strongly clustered according to the home country of the users. We, therefore, fix the number of clusters to 5 and perform clustering by the M-algorithm with three different objective functions (CND, IIW, and MIW). The results in Fig. 3 illustrate that utilizing CND clustering results is highly correlated with the grouping by country.

The other objective functions (IIW and MIW) were reported to achieve accurate clustering results both with the benchmark data and with the diagnosis clusters in [14, 19]. Especially IIW gained the best overall results and

Table 2 Proportion of links between country clusters (%)

Source	Target				
	Finland	Sweden	Norway	Denmark	Iceland
Finland	99.0	**0.6**	0.2	0.1	<0.0
Sweden	**1.1**	97.5	0.9	0.4	0.1
Norway	1.7	**7.2**	89.3	1.3	0.4
Denmark	1.8	**2.2**	1.0	94.8	0.2
Iceland	0.7	2.6	**3.6**	1.0	92.1

outperformed the other objective functions [19]. However, the goal of the application was to create balanced clusters, and the benchmark data was created accordingly. This is not the case here, as there is one much smaller cluster, Iceland. In clustering, it will be merged with the Denmark cluster when using the IIW function, and the Finland and Sweden clusters are split in an arbitrary manner. MIW also detects the communities quite poorly. Another difference is that the health data in is much more dense (average degree 139 for a network of 205 nodes) compared to the *NTN-2022* data [14], which is much sparser (average degree 3.4). For this reason, we use only the conductance objective function in the rest of this paper.

Five Clusters

The clustering results are visualized in Fig. 4, illustrating a clear correspondence to the country clusters. The cluster borders are quite clear, and very few users are clustered differently based on the interactional links than what their home country is. There are some weakly connected (almost isolated) components that are clustered differently. One visible is the small sub-cluster above the Iceland cluster, which contains mostly internal links within the cluster and very few links to other users in Iceland. In the case of conductance, it is put into the same cluster with Sweden. This appears to be an artifact of the algorithm.

The proportion of links between the different country clusters are summarized in Table 2. These numbers demonstrate that most links (>90%) are to users in the same country. This shows that users have strong connections within their home country and only weak connections with other users. Consequently, we can argue that there are five intrinsic clusters in the *NTN-2022* corresponding to the five Nordic countries. A possible explanation is the different languages used in each country.

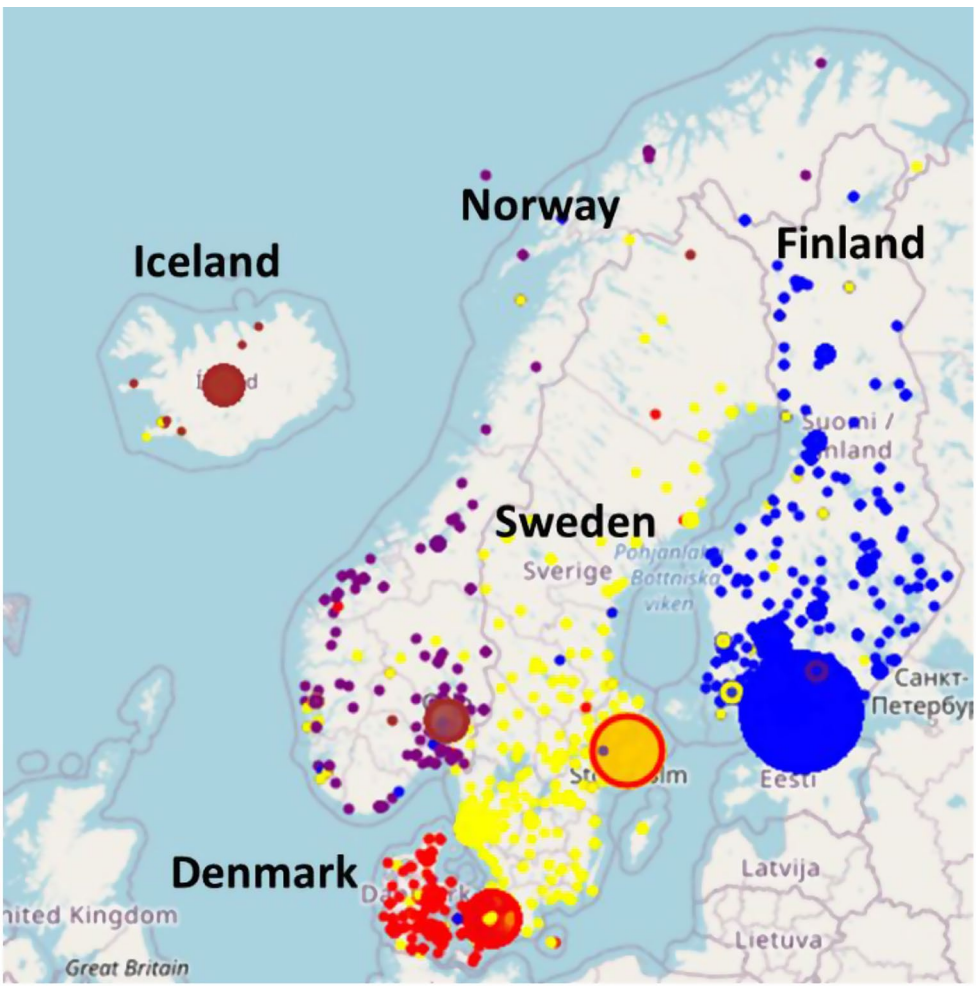

Fig. 5 Clustering results on map. Users are plotted at their home location and colored by the cluster it belongs to (Finland = blue, Sweden = yellow, Norway = purple, Denmark = red, Iceland = brown). Note: The geographical location of Iceland is artificially moved closer to Norway for making the figure more compact for easier analysis

Connections to other clusters are asymmetric. Finland is the most homogenous among the five countries, having 99% within cluster connections, possibly explained by its linguistic divergence from the other four (all Scandinavian) countries. Norway has the most links to other clusters (10.7%), of which most are to Sweden (7.2%). Sweden is the most linked from other clusters, probably explained by its central geographical location.

Visualizing the Clustering Results on Map

The users and their clusters are further visualized in Fig. 5 so that the users are plotted in their home locations and colored according to the country cluster they were assigned to. We did not detect any clear patterns in the user locations. One might expect that users in Finland who are clustered into the Sweden cluster might live in western Finland as most Swedish-speaking Finns live there. This is partly the case, but since there are so few users clustered outside their own home country, we cannot draw any strong conclusion.

Sixth Cluster

We investigated the data further by adding the sixth cluster to see if it would affect the result, see Fig. 6. The main observation is that the clustering result is no longer stable, and the result varies from one run to another because of randomness in the algorithm. Sometimes, it divides Denmark (left, also in Fig. 7), sometimes Sweden (middle), or it allocates the extra cluster to the small, almost isolated sub-cluster in Iceland (right). Sometimes, the extra cluster is merely a collection of borderline users that do not clearly belong to one country according to their interactional links. This instability indicates that the choice for the number of clusters (six) is inappropriate for the data [38].

We studied the situation further by dividing the Finland cluster into two sub-clusters (Fig. 8). This time, the result is stable, but there is only one real cluster. This becomes apparent when the two clusters are plotted separately (Fig. 9). Most users are in the bigger dense cluster, whereas the second cluster contains merely multiple disconnected subgraphs and nodes. These are mainly outliers that lack connections. We conclude that, based on the result here, it is unlikely that there would be any natural clusters present within any of the country clusters. The country clusters are strong, but users within one country cannot be divided further in a stable manner based on the interactional links alone.

Hashtag Analysis

We deepen the clustering based on interactional patterns in social networks with a content analysis by exploring the use of hashtags in the Tweets by the users in each home country cluster. Hashtags (#) are metadata tags used on social media platforms to allow users to label their posts by keywords [58]. Twitter supports hashtags by making it very easy for users to find tweets using a specific hashtag, which creates a discussion thread around any topic defined by a user. Other users can discover and join public conversations on particular topics, making hashtags a powerful tool for tracking social networks around user-generated content of specific themes.

For this purpose, we collected the most recent tweets from all users in the *NTN-2022*, with a maximum of 3,200 messages (the number of retrieved messages is limited by the Twitter API), and extracted the hashtags used in these tweets. We then calculated the number of unique hashtags for each country. The results in Table 3 display the percentage of tweets that include hashtags. The average number of hashtags per tweet is higher in the Finland cluster (0.72) than in the other countries (Sweden = 0.42, Norway = 0.46, Denmark = 0.50, Iceland = 0.45).

Top 10 hashtags of each country are then displayed in Table 4 (see supplementary material for hashtags

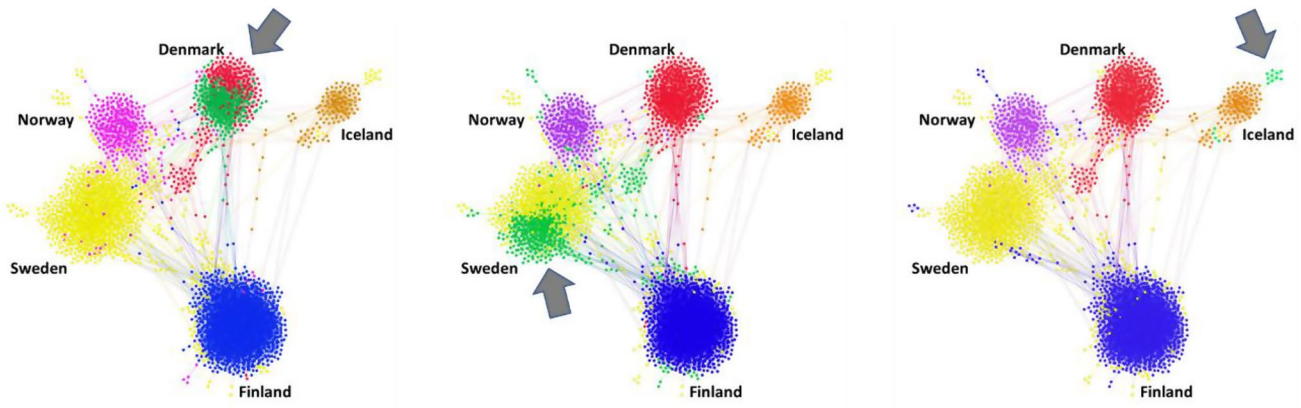

Fig. 6 Having six clusters does not lead to stable clustering results, and the additional cluster is highly sensitive to the initialization

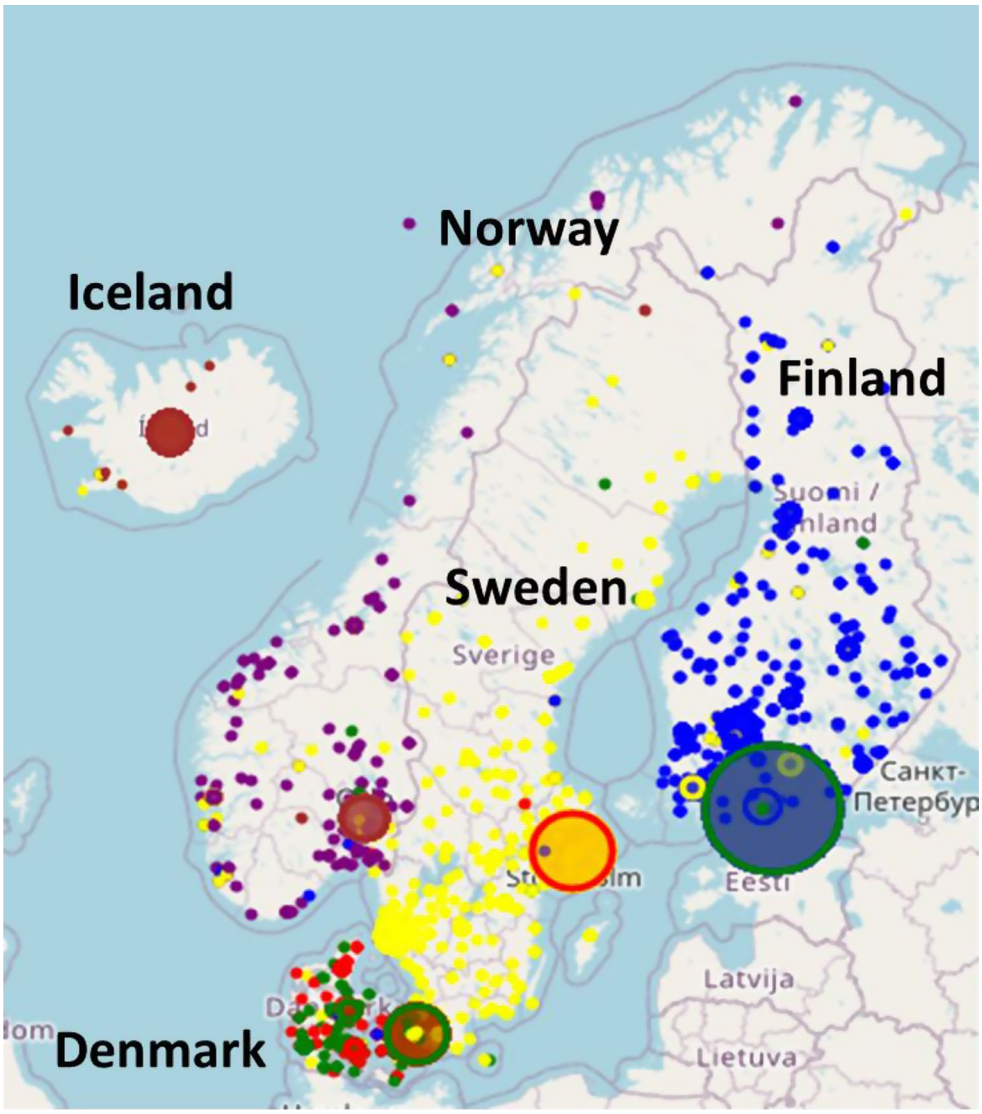

Fig. 7 Geographical distribution of the clusters in case when the additional cluster is allocated to Denmark. The statistics also showed high value of between these two clusters (14%) compared to the corresponding value of the Denmark cluster (5.2%)

descriptions). They are country-specific, and not even one hashtag appears in the top 30 lists of the other countries.

In Finland, the hashtags that appear most frequently are related to sports (7 occurrences) and location (2 occurrences), and the country-specific hashtag is ranked at #1. In Sweden, the most common hashtags include sports (3 occurrences), music (2 occurrences), politics (1 occurrence), location (1 occurrence), and other topics (2 occurrences), with the country hashtag ranked at #6.

Norway's most frequently used hashtags revolve around sports (9 occurrences), with the country hashtag ranked at #3. In Denmark, the prominent hashtags are politics (4 occurrences), sports (2 occurrences), awareness (2 occurrences), business (1 occurrence), and the country hashtag ranking is #9. Lastly, in Iceland, the hashtags that appear most frequently relate to sports (3 occurrences), tourism (3 occurrences), politics (1 occurrence), music (1 occurrence), other topics (1 occurrence), and the country hashtag holds the top ranking (#1).

By analyzing the most popular hashtags of each country separately, we can make a further observation that strengthens the conclusions based on clustering.

First, all countries have their country-specific hashtags in their corresponding top 10 lists (#finland, #sweden, #norway, #dksocial, #iceland). We observed that the smaller and less central countries among these five had the country hashtag higher in the rankings: Finland and Iceland (1st) and Norway (3rd). A possible interpretation is that the people in such countries have a stronger need to display their origin than people in bigger or more central countries. The largest country in the region in terms of population and the size of the economy are Sweden (6th) and Denmark (9th). Denmark is also more connected to continental Europe, which may further enhance this phenomenon.

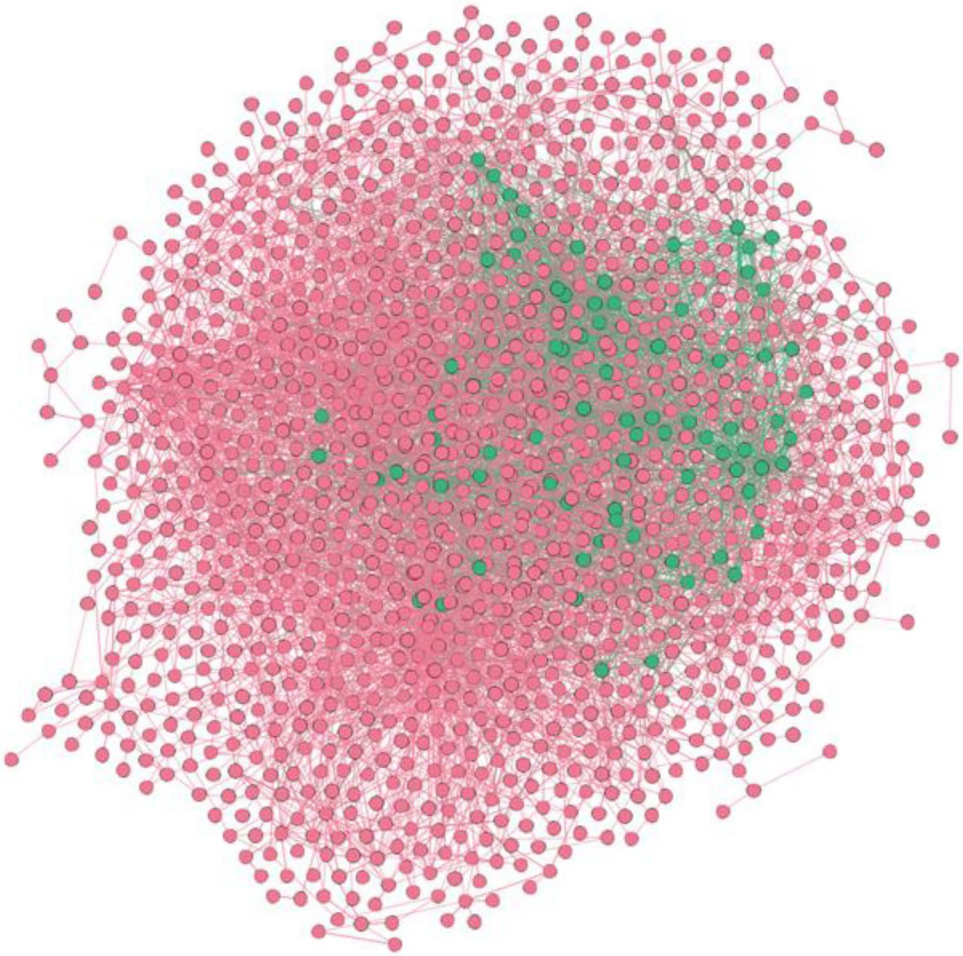

Fig. 8 Dividing Finland to two clusters seemed arbitrarily chosen without any natural explaining factor. The statistics also showed same value of between these two clusters (1%) compared to the corresponding value of the Finland cluster (1%)

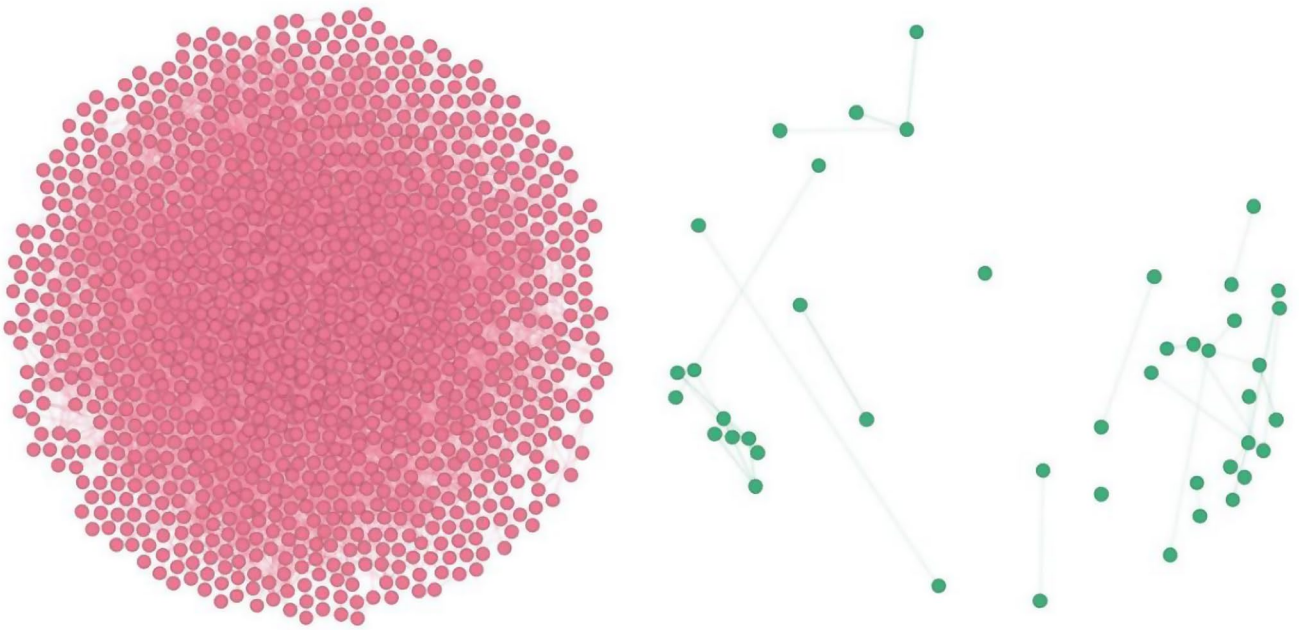

Fig. 9 The two sub-clusters inside Finland cluster consist mainly one large cluster. The second smaller cluster consists only isolated and weakly connected sub-clusters, i.e., outliers

Table 3 Statistics for the retrieved hashtags from up to 3,200 latest messages for *NTN-2022* users

	Tweets	Tweets with hashtags	Hashtags	Hashtags frequencies	Hashtags per tweet
Finland	2,392,135	28%	308,085	1,734,491	0.72
Sweden	6,137,063	22%	430,957	2,587,983	0.42
Norway	1,613,866	24%	169,009	748,077	0.46
Denmark	1,275,947	24%	138,010	641,646	0.50
Iceland	178,925	11%	26,756	81,231	0.45
Total	11,597,963	23%	1,072,817	5,793,428	0.50

Table 4 Top 10 most frequent hashtags for each country. For each country, hashtags are discerningly sorted based on their shares

	Finland	Sweden	Norway	Denmark	Iceland
1	finland	hundralappen	VierHBK	Dkpol	iceland
2	liiga	nowplaying	2pl	Sldk	fotboltinet
3	helsinki	twittpuck	Norway	dkmedier	12stig
4	ravit	Timraik	kolbotn	dkgreen	inspiredbyiceland
5	veikkausliiga	svpol	ffk1903	Obdk	fotbolti
6	huuhkajat	Sweden	2fx	Dkbiz	menntaspjall
7	sinipaidat	Ifkgbg	mufc	sundpol	lavacentre
8	tampere	Melfest	raufossfotball	Uddpol	kosningar
9	valioliiga	årebageri	bcfc	dksocial	skeidin
10	esportsfi	vitmagi	obosligae	Eudk	tiujardarnir

Table 5 Categorization of the most frequent hashtags

Finland	Sweden	Norway	Denmark	Iceland
Sports 7	Sports 3	Sports 9	Politics 4	Sports 3
Location 2	Music 2		Sports 2	Tourism 3
	Politics 1		Awareness 2	Politics 1
	Location 1		Business 1	Music 1
	Other 2			Other 1

Table 6 Hashtag similarity results (%)

	Finland	Sweden	Norway	Denmark	Iceland
Finland	-	7.2	**7.4**	6.8	2.7
Sweden	7.2	-	**8.0**	7.0	2.3
Norway	7.4	8.0	-	**10.0**	4.6
Denmark	6.8	7.0	**10.0**	-	5.0
Iceland	2.7	2.3	4.6	**5.0**	-

The content also demonstrated interesting differences. We further categorized the top 10 hashtags subjectively based on our understanding of their content, see Table 5. Sports-related hashtags were the most common. Norway had 9 hashtags (all except the country tag) about sports, and Finland had 7. They were mostly football (soccer) related, with the exceptions of ice hockey and horse race (Finland) and handball (Norway). The other countries had only 3 or 4 sports-related hashtags. Other common themes were politics (4 hashtags in Denmark), tourism (3 in Iceland), music (2 in Sweden) and awareness (2 in Denmark).

What we also examined is the similarity of the countries based on the overall use of hashtags. For this, we form the sets of all hashtags used in the same country. Jaccard Similarity Coefficient (JSC) [59] is then calculated as the number of common hashtags divided by the number of different hashtags in the two sets. The outcomes are shown in Table 6, where the maximum value of each row is emphasized.

Based on the results, Denmark and Norway share the most (10%) of all unique hashtags, whereas Iceland and Sweden have the least common hashtags (2.3%). The similarity in hashtags use does not align with the connectivity pattern between countries. Based on connectivity percentages in Table 2, the majority of the countries are mostly connected to Sweden, but hashtag use demonstrates the highest similarity to Norway.

Our results align with prior sociolinguistic studies highlighting strong national clustering patterns in multilingual digital spaces. Münch et al. [12] identified clear divisions between Italian and German Twitter communities based on language, similar to how our clusters correspond to national borders. Likewise, the authors in [6] found limited evidence of cross-national echo chambers in the Norwegian Twittersphere, which supports our observation of weak inter-country links. These parallels reinforce the conclusion that national identity and language are central in shaping social media interaction patterns, even in culturally and geographically close regions like the Nordics.

Limitations

We have focused on establishing a Nordic Twitter network based on the following/followee relations. One limitation of this approach is that it does not indicate the strength of the connections. An alternative approach would have been

to create a weighted network from the interactions (replies, mentions, and retweets) with a more fine-tuned network having potentially more information on the relations of the users. It would also allow different perspectives by considering the intensity and frequency. The chosen clustering algorithm would generalize to such network structure as well. This is a promising direction for future work.

A second limitation is the use of geo-location of users for the selection. It has the advantages of being more reliable and also having more expert users in the selection. However, it has a clear sub-sampling effect, which may become a limitation if we lack data from where to draw the sample. Fortunately, we had enough data.

We also excluded travelling users, i.e. those whose geo-location mismatches with some of the user's tweet location. This filtering was done to guarantee that we are selecting users only from the countries in question. As a side-effect, we might have lost some of the nuances in the data, which also made it easier for the clustering algorithm to detect country clusters. However, the filtering did not help to find any sub-clusters either. The method is a compromise of location accuracy and the richness of data.

A limitation of using hashtags as a selection criterion has also been noted in the literature [28–30]. However, we do not use hashtags for the selection, but only for the summarization of the content. Our focus is not to perform an extensive content analysis but to study whether there are natural clusters and, if yes, what they are. We found country clusters but no evidence of sub-clusters within a country. Future work could explore content-based clustering using embedding methods to extract deeper thematic insights from user-generated hashtags.

Other papers have reported intra-country clusters. However, it is possible to *create* some clusters by an algorithm even if the data (within a country) does not naturally divide into smaller clusters. In such cases, clustering just serves as a sub-sampling method. The smaller the clusters, the more differences there are in their content. However, we tried to *find* additional clusters but did not find evidence of them in data. A similar result was reported by the authors in [6], who did not find evidence of the echo chambers effect. While they might exist, a network built from the following/followee links is not able to reveal them.

Conclusions

We created a very large social network of Twitter users from the Nordic region. The data includes Nordic Twitter users who tweeted between 1 November 2016 and 31 December 2022. We then clustered the users according to their interactional links.

The main finding is that the clustering highly correlates with the home country of the users with only minor differences. Finland had the highest share (99%) of interaction connections within the same country, and Sweden had the smallest (89%). The result is surprising considering that four of the five countries share similar (typologically Germanic) languages and similar cultures [26]–[27]. However, their topics on Twitter are very country-specific, and most friends are in the same country.

We also added a sixth cluster, but the result was either unstable or, in the case of Finland, the algorithm just created an additional location outlier cluster. It implies that there is no natural additional cluster within any of the countries based on the interactional links. Further analysis of the hashtag data within country clusters indicated a clear pattern: every country had mainly their own topics. The results also showed some country specific behavior in the selection of hashtags. For example, Finland and Iceland had the country name as the #1 hashtag. Another example is that sports themes were highly popular in Norway and Finland but less so in Sweden and Denmark.

Despite shared geography and cultural similarities, social media users in the Nordic region cluster strongly along national lines, indicating that digital interactions continue to reflect offline national identities. This insight benefits policymakers and sociologists exploring digital cohesion and platform designers seeking to enhance cross-border or multilingual engagement.

Further research should focus more on combining network-related information with more extensive user-generated textual content, including detecting trends and how the topics evolve over time. The data would also allow comparison of more profound linguistic differences that vary over time and across geographical locations. Similar to [60], it would be possible to examine linguistic factors associated with English usage in non-native English-speaking countries by considering the interactional patterns, topological properties, and connections among Nordic Twitter users.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42979-025-04353-y.

Acknowledgements 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 Application (DISA) at Linnaeus University.

Funding Open access funding provided by University of Eastern Finland (including Kuopio University Hospital).

Data Availability The graph datasets documented in Fig. 2 are published in https://github.com/uef-machine-learning/NTN-2022.

Code Availability The algorithms' source code is available in: https://github.com/uef-machine-learning/gclu.

Declarations

Competing Interests All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Arazzi M, Ferretti M, Nicolazzo S, Nocera A. The role of social media on the evolution of companies: a Twitter analysis of streaming service providers. Online Social Networks Media. 2023;36. https://doi.org/10.1016/j.osnem.2023.100251.
2. Yao Q, Li RYM, Song L, Crabbe MJC. Construction safety knowledge sharing on twitter: A social network analysis. Saf Sci. 2021;143:105411. https://doi.org/10.1016/j.ssci.2021.105411.
3. Motamedi R, Jamshidi S, Rejaie R, Willinger W. Examining the evolution of the Twitter elite network. Social Netw Anal Min. 2020;10(1). https://doi.org/10.1007/s13278-019-0612-8.
4. Laitinen M, Fatemi M. Data-intensive sociolinguistics using social media. Ann Academiae Scientiarum Fennica. 2023;2023(2):38–61. https://doi.org/10.57048/aasf.136177.
5. Bruns A, Burgess J, Highfield T. A 'big data' approach to mapping the Australian Twittersphere. In: Arthur PL, Bode K, ed. Advancing digital humanities. London: Palgrave Macmillan; 2014. pp. 113–29. https://doi.org/10.1057/9781137337016_8.
6. Bruns A, Enli G. The Norwegian twittersphere: structure and dynamics. Nordicom Rev. 2018;39(1):129–48. https://doi.org/10.2478/nor-2018-0006.
7. Bruns A, Moon B. One day in the life of a National Twittersphere. Nordicom Rev. 2019;40(s1):11–30. https://doi.org/10.2478/nor-2019-0011.
8. Geenen DV, Schaefer MT, Boeschoten T, Hekman E, Bakker P, Moons J. (2016, October). Mining One Week of Twitter. Mapping networked publics in the dutch twittersphere, The 17 annual conference of the association of internet researchers. Berlin, Germany. https://api.semanticscholar.org/CorpusID:158218716
9. Kwak H, Lee C, Park H, Moon S. (2010). What is Twitter, a social network or a news media? In proceedings of the 19th international conference on World wide web (WWW '10), Association for Computing Machinery, New York, NY, USA, 591–600. https://doi.org/10.1145/1772690.1772751
10. Münch FV, Rossi L. (2020). Bootstrapping Follow Networks of Influential Twitter Accounts, IC2S2. https://vimeo.com/431470176
11. Münch FV, Rossi L. A Tale of two twitters? Identifying bridges between Language based twitters?pherees. AoIR Sel Papers Internet Res. 2020. https://doi.org/10.5210/spir.v2020i0.11283.
12. Münch FV, Thies B, Puschmann C, Bruns A. Walking through twitter: sampling a Language based follow network of influential Twitter accounts. Social Media + Soc. 2021;7(1). https://doi.org/10.1177/2056305120984475.
13. Mishra N, Schreiber R, Stanton I, Tarjan RE. Clustering social networks. In: Bonato A, Chung FRK, ed. Algorithms and models for the Web-Graph. WAW 2007. Lecture Notes in Computer Science. Volume 4863. Berlin, Heidelberg: Springer; 2007. https://doi.org/10.1007/978-3-540-77004-6_5.
14. Fränti P, Sieranoja S, Wikström K, Laatikainen T. Clustering diagnoses from 58 M patient visits in Finland between 2015 and 2018. JMIR Med Inf. 2022;10(5):e35422. https://doi.org/10.2196/35422.
15. Ramasubbareddy S, Srinivas TAS, Govinda K, Manivannan SS. Comparative study of clustering techniques in market segmentation. In: Saini H, Sayal R, Buyya R, Aliseri G, ed. Innovations in computer science and engineering. Singapore: Springer; 2020. pp. 117–25. https://doi.org/10.1007/978-981-15-2043-3_15.
16. Almannaa MH, Elhenawy M, Rakha HA. A novel supervised clustering algorithm for transportation system applications. IEEE Trans Intell Transp Syst. 2020;21(1):222–32. https://doi.org/10.1109/TITS.2018.2890588.
17. Kernighan BW, Lin S. An efficient heuristic procedure for partitioning graphs. Bell Syst Tech J. 1970;49(2):291–307. https://doi.org/10.1002/j.1538-7305.1970.tb01770.x.
18. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
19. Sieranoja S, Fränti P. Adapting k-means for graph clustering. Knowl Inf Syst. 2021;64:115–42. https://doi.org/10.1007/s10115-021-01623-y.
20. Bruns A, Moon B, Münch F, Sadkowsky T. The Australian Twittersphere in 2016: mapping the follower/followee network. Social Media + Soc. 2017;3(4). https://doi.org/10.1177/2056305117748162.
21. Rosvall M, Bergstrom CT. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PLoS ONE. 2011;6(4):e18209. https://doi.org/10.1371/journal.pone.0018209.
22. Peixoto TP. Parsimonious module inference in large networks. Phys Rev Lett. 2013;110(14):148701. https://doi.org/10.1103/PhysRevLett.110.148701.
23. Peixoto TP. Descriptive vs. Inferential community detection in networks: pitfalls, Myths and Half-Truths. Cambridge: Cambridge University Press; 2023. https://doi.org/10.1017/9781009118897.
24. Laitinen M, Lundberg J, Levin M, Martins RM. (2018). The Nordic Tweet Stream: a dynamic real-time monitor corpus of big and rich language data. DHN 2018 Digital humanities in the nordic countries 3rd conference: proceedings of the digital humanities in the nordic countries 3rd conference Helsinki, Finland, pp. 349–362. https://erepo.uef.fi/handle/123456789/6697
25. Zheng X, Han J, Sun A. A survey of location prediction on Twitter. IEEE Trans Knowl Data Eng. 2018;30(9):1652–71. https://doi.org/10.1109/TKDE.2018.2807840.
26. McArthur T. World english, Euro-English, nordic english? Engl Today. 2003;19(1):54–8. https://doi.org/10.1017/S026607840300107X.

27. Kristiansen T, Sandøy H. Introduction. The linguistic consequences of globalization: the nordic laboratory. Int J Sociol Lang. 2010;204:1–7. https://doi.org/10.1515/ijsl.2010.027.
28. Bruns A, Burgess J. (2015). Twitter hashtags from ad hoc to calculated publics. Hashtag Publics: Power politics discursive networks, 13–28.
29. Carpenter J, Tani T, Morrison S, Keane J. Exploring the landscape of educator professional activity on twitter: an analysis of 16 education-related Twitter hashtags. Prof Dev Educ. 2022;48(5):784–805. https://doi.org/10.1080/19415257.2020.1752287.
30. Lattimer TA, Ophir Y. Oppression by omission: an analysis of the #whereistheinterpreter hashtag campaign around COVID-19 on Twitter. Media Cult Soc. 2023;45(4):769–84. https://doi.org/10.1177/01634437221135977.
31. Prasetyo PK, Achananuparp P, Lim EP. (2016, January). On analyzing geotagged tweets for location-based patterns. In Proceedings of the 17th International Conference on Distributed Computing and Networking (ICDCN '16). Association for Computing Machinery, New York, NY, USA, Article 45, 1–6. https://doi.org/10.1145/2833312.2849571
32. Graham M, Hale SA, Gaffney D. Where in the world are you? Geolocation and Language identification in Twitter. Prof Geogr. 2014;66(4):568–78. https://doi.org/10.1080/00330124.2014.907699.
33. Milroy J, Milroy L. Linguistic change, social network and speaker innovation. J Linguist. 1985;21(2):339–84. http://www.jstor.org/stable/4175792.
34. Gonçalves B, Perra N, Vespignani A. Modeling users' activity on Twitter networks: validation of dunbar's number. PLoS ONE. 2011;6(8):1–5. https://doi.org/10.1371/journal.pone.0022656.
35. Laitinen M, Fatemi M, Lundberg J. Size matters: digital social networks and language change. Front Artif Intell. 2020;3:46. https://doi.org/10.3389/frai.2020.00046.
36. Bastian M, Heymann S, Jacomy M. (2009). Gephi: An Open Source software for exploring and manipulating networks, proceedings of the third international AAAI conference on weblogs and social media. San José, Unite States. https://doi.org/10.1609/icwsm.v3i1.13937
37. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PLoS ONE. 2014;9(6):e98679. https://doi.org/10.1371/journal.pone.0098679.
38. Rezaei M, Fränti P. Can the number of clusters be determined by external indices? IEEE Access. 2020;8:89239–57. https://doi.org/10.1109/ACCESS.2020.2993295.
39. Schaeffer SE. Graph clustering. Comput Sci Rev. 2007;1(1):27–64. https://doi.org/10.1016/j.cosrev.2007.05.001.
40. Kannan R, Vempala S, Vetta A. On clusterings: good, bad and spectral. J ACM. 2004;51(3):497–515. https://doi.org/10.1145/990308.990313.
41. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci. 2002;99(12):7821–6. https://doi.org/10.1073/pnas.122653799.
42. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. (2004). Defining and identifying communities in networks. Proceedings of the National Academy of Sciences, 101(9), pp. 2658–2663. https://doi.org/10.1073/pnas.0400054101
43. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E. 2006;74(3):036104. https://doi.org/10.1103/PhysRevE.74.036104.
44. Whang JJ, Gleich DF, Dhillon IS. Overlapping community detection using Neighborhood-Inflated seed expansion. IEEE Trans Knowl Data Eng. 2015;28:1272–84. https://api.semanticscholar.org/CorpusID:11934509.
45. Chartrand G, Erdös P, Oellermann OR. How to define an irregular graph. Coll Math J. 1988;19(1):36–42. https://doi.org/10.1080/07468342.1988.11973088.
46. Karypis G, Kumar V. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal Sci Computing. 1998;20(1):359–92. https://doi.org/10.1137/S1064827595287997.
47. Rozemberczki B, Davies R, Sarkar R, Sutton C. (2020) GEMSEC: graph embedding with self clustering, In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19). Association for Computing Machinery, New York, NY, USA, 65–72. https://doi.org/10.1145/3341161.3342890
48. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. 2004;69(2):026113. https://doi.org/10.1103/PhysRevE.69.026113.
49. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations. Proc 20th ACM SIGKDD Int Conf Knowl Discovery Data Min. 2014;701–710. https://doi.org/10.1145/2623330.2623732.
50. Hamilton WL, Ying R, Leskovec J. (2017). Inductive representation learning on large graphs. proceedings of the 31st international conference on neural information processing systems, 1025–1035.
51. Kirkley A, Newman MEJ. Representative community divisions of networks. Communication Phys. 2022;5(1). https://doi.org/10.1038/s42005-022-00816-3.
52. Wang L, Zhang Z, Dunson D. Common and individual structure of brain networks. Annals Appl Stat. 2019;13(1):85–112. https://doi.org/10.1214/18-AOAS1193.
53. Young JG, Cantwell GT, Newman MEJ. Bayesian inference of network structure from unreliable data. J Complex Networks. 2020;8(6):cnaa046. https://doi.org/10.1093/comnet/cnaa046.
54. Young JG, Kirkley A, Newman MEJ. Clustering of heterogeneous populations of networks. Phys Rev E. 2022;105(1):014312. https://doi.org/10.1103/PhysRevE.105.014312.
55. Fränti P, Sieranoja S. How much k-means can be improved by using better initialization and repeats? Pattern Recogn. 2019;93:95–112. https://doi.org/10.1016/j.patcog.2019.04.014.
56. Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell. 2000;22(8):888–905. https://doi.org/10.1109/34.868688.
57. Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth. Knowl Inf Syst. 2015;42:181–213. https://doi.org/10.1007/s10115-013-0693-z.
58. Fatemi M, Kucher K, Laitinen M, Fränti P. Selfsimilarity of Twitter users. 2021 Swed Workshop Data Sci (SweDS). 2021;1–7. https://doi.org/10.1109/SweDS53855.2021.9638288.
59. Thada V, Jaglan V. Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. Int J Innovations Eng Technol. 2013;2(4):202–5.
60. Taipale I, Laitinen M. Individual sensitivity to change in the lingua Franca use of english. Front Communication. 2022. https://doi.org/10.3389/fcomm.2021.737017. 6.