PUBLICATIONS OF
THE UNIVERSITY OF EASTERN FINLAND

Dissertations in Science, Forestry and Technology

MASOUD FATEMI

COMPUTATIONAL METHODS FOR ANALYSING TWITTER-BASED SOCIAL NETWORKS

PUBLICATIONS OF THE UNIVERSITY OF EASTERN FINLAND
DISSERTATIONS IN SCIENCE, FORESTRY AND TECHNOLOGY

N:o 143

Masoud Fatemi

COMPUTATIONAL METHODS FOR ANALYSING TWITTER-BASED SOCIAL NETWORKS

ACADEMIC DISSERTATION

To be presented by the permission of the Faculty of Science, Forestry and Technology for public examination in the Auditorium M103 in Metria Building at the University of Eastern Finland, Joensuu, on 16 June 2026 at 12 o'clock noon.

University of Eastern Finland
School of Computing
Joensuu 2026

PunaMusta Oy
Joensuu, 2026
Editor: Pertti Pasanen, Nina Hakulinen, Raine Kortet, Matti Tedre, Pasi Vahimaa and Timo Lähivaara

ISBN: 978-952-61-6012-2 (print)
ISSNL: 2954-131X
ISSN: 2954-131X
ISBN: 978-952-61-6013-9 (pdf)
ISSNL: 2954-131X
ISSN: 2954-1484

Author's address:
University of Eastern Finland
School of Computing
P.O.Box 111
80101 JOENSUU
FINLAND
email: masoud.fatemi@uef.fi

Supervisors:
Professor Pasi Fränti
University of Eastern Finland
School of Computing
P.O.Box 111
80101 JOENSUU
FINLAND
email: pasi.franti@uef.fi

Professor Mikko Laitinen
University of Eastern Finland
School of Humanities
P.O.Box 111
80101 JOENSUU
FINLAND
email: mikko.laitinen@uef.fi

Reviewers:
Professor Tapio Pahikkala
University of Turku
Department of Computing
20014 TURKU
FINLAND
email: tapio.pahikkala@utu.fi

Professor Jari Saramäki
Aalto University
Department of Computer Science
P.O. Box 11000
00076 ESPOO
FINLAND
email: jari.saramaki@aalto.fi

Opponent:
Professor Tuomo Hiippala
University of Helsinki
Department of Languages
P.O.Box 3
00014 HELSINKI
FINLAND
email: tuomo.hiippala@helsinki.fi

Masoud Fatemi
Computational methods for analysing Twitter-based social networks
Joensuu: University of Eastern Finland, 2026
Publications of the University of Eastern Finland
Dissertations in Science, Forestry and Technology: 143

ABSTRACT

Online social networks have become the primary arenas for social communication and behavior. Analyzing these environments requires scalable methods that go beyond traditional, small-scale manual approaches. In this thesis, we focus on Twitter data and computational approaches to model social behaviour, similarity, communication, and tie strength at a large scale.

Classical sociological concepts such as the distinction between weak and strong ties and the notion of user similarity remain analytically valuable, yet they are difficult to operationalize consistently in large datasets. Existing approaches to measuring tie strength are often limited to edge- or node-level metrics, depend on ad hoc thresholding, and are grounded in ethnographic observations that do not easily generalize to large-scale network data. To address these challenges, this thesis develops a scalable computational framework for modelling Twitter-based social networks and introduces a network-level, multidimensional method for quantifying and comparing tie strength within ego networks.

This thesis draws on several Twitter datasets, including an ego network collection, a survey-based dataset paired with Twitter network streams for measuring user similarity, a geo-labeled Nordic Twittersphere network, and a multi-region corpus comprising Nordic, UK, US, and Australian data for tie strength analysis. All datasets were built using the now-discontinued Twitter Academic API. Across these datasets, the thesis conducts diffusion analysis and evaluates alternative similarity signals derived from interactions, activity patterns, and user-generated content. It further applies graph clustering methods to detect communities and introduces a Network Strength Index (NSI), a composite measure built from eight indicators, to quantify tie strength at the ego network level. Finally, ego networks are clustered into ordered tie strength groups using the most informative NSI measures, enabling large-scale comparisons of network structures across geographic regions and genders.

The key findings of this thesis are as follows. First, the effect of weak-tie and strong-tie networks on innovation diffusion depends on network size. Examining language change as a case of diffusion, we demonstrate that once an ego network reaches approximately 120 nodes, the behavioral distinction between weak-tie and strong-tie environments largely disappears: both network types exhibit similar diffusion dynamics. Second, we find that among several approaches to measuring user similarity on Twitter, interaction-based similarity aligns most closely with human judgment and outperforms similarity measures based on activity patterns or content. Across all methods, however, similarity estimation becomes less accurate as network size increases. Third, clustering the Nordic Twittersphere reveals that its community structure follows national boundaries: countries cluster distinctly, with no notable cross-country superclusters and no meaningful sub-national clusters. Fourth, our comparison of eight computational measures for estimating tie

strength indicates that interaction strength (IS), social similarity (SS), and the outliers (OUT) are the most informative and robust indicators. These measures remain stable across networks of varying sizes and degrees, making them the strongest candidates for large-scale tie strength analysis. Fifth, using the most informative measures selected through the NSI framework, ego networks can be clustered into ordered tie strength categories. This enables systematic comparison across populations. Applying this method to the multiregional corpus reveals consistent regional patterns: Nordic networks have the weakest overall tie strength, while Australian networks exhibit the strongest share of strong ties.

This thesis builds a bridge between traditional social theories often grounded in small-scale or ethnographic observations and scalable computational analysis. It does so by developing a framework for comparing complete social structures, such as ego networks, across large-scale Twitter datasets. By operationalizing classic concepts in a way that generalizes beyond local observations, the thesis provides a methodological foundation for future research on innovation diffusion, community structure, and regional or demographic variation in online social networks.

Keywords: *Online social networks, network modelling, tie strength, user similarity, ego networks, graph clustering, innovation diffusion*

ACKNOWLEDGEMENTS

Once upon a time, there was a boy growing up in a city on the edge of the desert called Jajarm. He dreamed of going "kharej" (i.e., abroad) to pursue his studies. I can now say that he did. Many have done so before, and many will do so after, but he followed his own path and carries his own story. From playing football in the streets under the blazing 30-degree sun with childhood friends to ice swimming during –30-degree Finnish winters, the journey has been extraordinary. Yet this achievement is not mine alone. I would not be here without the support of several remarkable people who shaped this path and made it possible.

I would like to express my gratitude to my supervisor, Prof. Pasi Fränti, for his guidance and perspective throughout this journey. His mentorship has shaped not only how I approach research but also how I think about collaboration, mentorship, and the kind of academic I aspire to become. I am deeply grateful to my other supervisor, Prof. Mikko Laitinen, for his unwavering support throughout my doctoral studies. His academic guidance, critical insights, and encouragement have been invaluable at every stage, and his trust and belief in my abilities have given me the confidence to grow as a researcher. Beyond his academic guidance, I am deeply grateful for his kindness, generosity, and integrity. I also sincerely acknowledge his financial support, which made it possible for me to pursue and complete my doctoral research.

I sincerely thank my pre-examiners, Prof. Jari Saramäki and Prof. Tapio Pahikkala, for their constructive comments, which enhanced the quality of this thesis.

I am grateful to the University of Eastern Finland and the School of Computing for providing a supportive academic environment and the resources necessary to complete my PhD. In addition, I am grateful to the Center for Data Intensive Sciences and Applications (DISA) at Linnaeus University for their financial support during the early years of my doctoral studies.

My thanks also go to Dr. Paula Rautionaho, Radu Mariescu-Istodor, and Dr. Kostiantyn Kucher for their guidance and for the many insightful, if occasional, conversations that left a lasting impression. I am grateful to my colleagues: Jimi, Abigail, Gulraiz, Irene, and many others, who have been more like friends than colleagues. Their companionship, support, and shared experiences made this journey not only manageable but truly enjoyable.

My deepest gratitude goes to my parents, Zahra and Jalal, whom I have missed dearly throughout these years. I still hear your voices in my mind: the quiet encouragement, the small reminders, the everyday care that shaped who I am. Everything I have achieved rests on the sacrifices you made long before I understood their weight. I am also grateful to my in-laws, Simin and Teymour, whose kindness and unconditional love warm my heart. Additionally, I would like to thank my brothers and sister, whom I have not been able to hug for far too long; those moments of making fun of one another and wrestling together are deeply missed. I miss you all.

I am equally thankful for my second family here in Joensuu: Henrietta, Joona, Mehrdad, Piotr, Arash, Laís, Inka, Paula, Ana, Nelli, Janne, Gaëlle, Arturo, Mia, and many others. We have traveled together, laughed together, cried together, and gone a little crazy together. Your friendship, support, and presence have meant more to me than words can express. I hope I can hold you close for the rest of my life. I once heard that "a friend is the brother you get to choose." Mehdi, you have truly been that brother to me. Over the past 13 years, we have traveled the world together and created countless unforgettable memories, and I am certain there are many more

ahead of us. Thank you for being there from the very beginning and for helping make this journey possible.

I have had the chance to meet and get to know many amazing people in my life, but you, Nazanin, were different from the very beginning. Thinking about you, and about the love I carry in my heart for you, a love I am sure you cannot fully understand, makes my hands tremble, and my eyes fill with tears. Thank you for always being there, for the sweetness that makes even my bitter days bearable, and for everything you are. I love you. You are my reason.

As I write this chapter, my fellow Iranians in my beloved homeland are enduring a profoundly difficult period in our history, marked by suffering and oppression. I dedicate this dissertation to all fathers and mothers, and to the fearless souls who laid down their lives for a better tomorrow.

Joensuu, April 27, 2026

Masoud Fatemi

LIST OF PUBLICATIONS

This thesis consists of the present review of the author's work in the field of social network modelling and the following selection of the author's publications:

I Laitinen, M., **Fatemi, M.**, & Lundberg, J. (2020). Size matters: Digital social networks and language change. *Frontiers in Artificial Intelligence*, 3. https://doi.org/10.3389/frai.2020.00046

II **Fatemi, M.**, Kucher, K., Laitinen, M., & Fränti, P. (2021). Self-similarity of Twitter users. *2021 Swedish Workshop on Data Science*, 1–7. https://doi.org/10.1109/SweDS53855.2021.9638288

III **Fatemi, M.**, Sieranoja, S., Laitinen, M., & Fränti, P. (2025). Detecting connectivity patterns in Nordic Twittersphere by cluster analysis. *SN Computer Science*, 6(7), 815. https://doi.org/10.1007/s42979-025-04353-y

IV **Fatemi, M.**, Laitinen, M., & Fränti, P. (2026). Computer-mediated communication and networks: Quantifying ego network strength. In Special Issue on Computer-Mediated Communication Corpora. *Language@Internet*.[Accepted for publication]

V **Fatemi, M.**, Laitinen, M., & Fränti, P. (2025). Clustering digital ego networks by tie strength: A scalable, platform-independent method. *The 20th International Conference on Intelligent Systems and Knowledge Engineering*. Shunde, China.

Throughout the thesis, these papers will be referred to by Roman numerals and are included at the end of this thesis.

AUTHOR'S CONTRIBUTION

The publications selected in this dissertation are original research papers on social network modelling.

The ideas presented in papers **I**–**III** originated from the co-authors. In paper **I**, the author developed the computational method, constructed the dataset, and contributed to the analysis and visualization. The paper **I** was jointly written with the co-authors. In papers **II** and **III**, the author implemented the methods, conducted all experiments, and produced all visualizations. These papers were also co-authored, with the author responsible for the majority of the writing.

The ideas for papers **IV** and **V** originated from the author. The author implemented the methods, conducted all experiments, and created all visualizations. Both papers were written in collaboration with co-authors, with the author responsible for the majority of the writing.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

4D	4-Dimensional
ABC	Average of betweenness centrality
ACC	Asymmetry closeness centrality
API	Application Programming Interface
CMC	Computer-Mediated Communication
CSC	Center for Science
DENS	Density
DSN	Digital Social Network
IS	Interaction Strength
MAD	Median Absolute Deviation
NLP	Natural Language Processing
NSI	Network Strength Index
NTS	Nordic Tweet Stream
OSN	Online Social Network
OUT	Outliers
RBC	Range of betweenness centrality
RIS	Relative Interaction Strength
SN	Social Network
SS	Social Similarity
SSE	Sum of Squared Error

1 Introduction

A *social network* (SN) refers to a set of actors (*nodes*) and the relationships (*ties*) that connect them. Nodes can represent individuals, entities, and organizations [1]. Ties can represent different kinds of relationships, such as friendships, professional connections, or family relationships [2,3].

Social networks exist offline and online. Offline social networks are based on physical, face-to-face interactions among individuals or organizations and are often shaped by geographical proximity [4]. These networks form through direct contact in everyday contexts such as family, workplace, and school [4].

Online social networks are formed through digital platforms [5]. In contrast to offline social networks, where interactions are based on physical co-presence, online social networks support interactions within the platform and via the internet. In other words, interactions in these networks are computer-mediated and not limited by geographical distance [5]. Social media networks such as Facebook, Instagram, Twitter/X, Bluesky, and online forums such as Reddit fall into this category.

Social networks, whether online or offline, are about relationships, not just individuals. They are the fundamental part of human life and the basic structures through which individuals connect, interact, exchange information, and form communities [6]. Analyzing these social structures helps to understand how nodes interact within a network, providing insights into community dynamics and the information flow [3].

In this thesis, the focus is on online social networks, specifically those from social media. Over the past two decades, these networks have expanded rapidly, becoming so deeply intertwined with everyday life [7]. They influence what we know, whom we trust, and how we behave, as well as making daily life easier by helping us find jobs, share news, or learn a new language or skill [8].

Studying how online social networks function now matters across many fields, from linguistics and sociology to computer science and public policy [9,10]. What people read, the opinions they form, and even how online communities grow are all shaped by online networks. These systems give rise to many familiar online dynamics, including the rapid spread of misinformation, group polarization, and the emergence of new words or expressions across platforms [11–13].

Recent studies also in the digital humanities and social sciences describe a broader shift, the network turn, in which networks have moved beyond their origins as a technical concept in computer science [14]. They have become a new framework for thinking across disciplines in society, culture, and communication [14]. This shift highlights the need to approach, for instance, sociolinguistic phenomena as fundamentally network-dependent: what matters is not only *what* is written, but also *to whom* it is written and how individuals are connected [14]. At the same time, a network model is an abstraction of the real world; while highlighting some relations, it hides others. Interpretation and methodological choices, therefore, play a crucial role, especially as large-scale network data and computational techniques are increasingly used [14]. Accordingly, researchers must remain critically aware of what network methods enable and how they can be utilized. This perspective

aligns with the aim of this thesis, adopting computational network modelling to analyze Twitter-based social networks while keeping these interpretive and societal implications in view.

Although online social networks have a substantial impact on human behavior, it is still challenging to study these networks and analyze the interactions within them in practice. One of the key challenges is connecting the long-standing social theories, which are often based on small ethnographic observations, to modern computation methods that can handle large and complex datasets [9, 10]. This thesis aims to establish this connection via utilizing data-driven modelling and tries to facilitate a better understanding of how people behave and interact in today's networked data.

1.1 OVERVIEW OF THE DISSERTATION

This thesis focuses on computational methods for analyzing large-scale online social networks derived from Twitter (currently known as X). It is a compilation thesis based on research conducted between 2019 and 2025. The work brings together a series of papers, each addressing different aspects of how online networks can be modelled, clustered, and analyzed at scale. Collectively, these papers introduce new computational approaches and evaluate traditional social network theories, such as the weak-tie hypothesis, in large digital environments.

The structure of the dissertation is as follows. In Chapter 2, we concentrate on why social networks matter and explain why modelling them is not a straightforward task. We also explore the main research questions that this thesis aims to answer. Chapter 3 introduces the main social media platform and its characteristics, which served as the primary data source for this thesis. Chapter 4 presents methods for modeling social networks, focusing on both traditional and computational approaches as well as a proposed multidimensional approach. Chapter 5 explores five applications of network modeling. It examines how connections are used in language diffusion, user similarity analysis, community detection, and tie-strength measurement. Chapter 6 summarizes the main contributions of the included papers and outlines directions for future research. Finally, Chapter 7 concludes the dissertation.

2 The role of social networks

There is a well-established theory in social network studies that categorizes networks based on the strength of their ties into *weak-tie* and *strong-tie* networks [15]. Weak ties, such as those among distant colleagues or acquaintances, are valuable for accessing novel information and opportunities, as they connect individuals to diverse social circles. In contrast, strong ties, such as ties among close friends or family members, provide emotional support and trust. However, because individuals within tightly knit groups tend to share similar experiences and knowledge, strong ties may lead to redundant information [15].

With the expansion of online platforms, researchers have become increasingly interested in applying social network theory to digital environments [5]. In addition, as social media usage continues to grow, these platforms have turned into vast repositories of user-generated content [16]. In other words, repositories that collect data from social media provide rich datasets in terms of interactional information and metadata for analysis [16, 17]. Consequently, researchers are increasingly turning to applying computational methods to analyze these large-scale datasets, moving beyond traditional, small-scale approaches [I, III, IV, V, 18].

Analyzing online social networks and their massive user-generated content dataset has applications across various domains, such as information diffusion [19], language change [I, 20, 21], the formation of location-based communities [III, 22], and connecting individuals based on shared characteristics [II, 23].

Recent research has begun to question the extent to which Granovetter's theory [15] on the strength of weak ties applies in digital environments [I, 18]. In particular, how weak ties influence language change in online settings [I, 21, 23, 24]. Additionally, recent literature has examined how network structures and the strength of ties between users affect language use on social media platforms [20, 21].

Tie strength can also influence information sharing patterns within a network. Social networks enable individuals to disseminate content to a wide range of audiences [23]. Research indicates that the likelihood of users sharing content increases when their friends have already shared the same content [23]. Previous studies suggest that weak ties expose users to non-redundant and novel information. In contrast, strong ties might be more influential but often lead to the circulation of redundant information [25, 26]. Accordingly, recent research has investigated the extent to which exposure to social signals influences information-sharing patterns on platforms such as Twitter and Facebook [20, 23].

Online social networks are also influential in language change. They operate as the pathways through which linguistic innovations spread [I, 20, 21]. Del Tredici and Fernández [21] argued that new words and language patterns often emerge from central community members with relatively weak ties. These innovations are then adopted and propagated by users with strong ties within more tightly connected subgroups [21]. This highlights the fundamental role of social networks in language change within communities [27, 28].

Another area where social networks play a significant role is forming location-based communities by connecting users who reside in the same geographic area [III,

22]. Our results based on Twitter data demonstrate that users tend to cluster strongly according to their home country [**III**]. In addition, although it is easy to connect globally, there is little interaction among people in nearby but different countries [**III**].

Homophily refers to the tendency of people with similar traits to form connections, and social networks play an important role in it [23]. There is also another form of similarity called *perceived similarity*, which refers to how similar an individual seems to another person based on personal opinions [**II**]. Social networks impact perceived similarity and how users consider themselves similar to others and establish connections based on shared activities [**II**].

In this thesis and through modeling large Twitter networks, we study these dynamics. Each paper (i.e., **I**, **II**, **III**, **IV**, and **V**) focuses on a specific aspect, ranging from language diffusion and user similarity to community structure and tie strength, providing an integrated view of how social connections operate in digital environments.

In summary, the role of social networks across various domains highlights the importance of analyzing and understanding them. However, several key challenges arise when modeling these networks [III, 20, 28–32]. The following section outlines the key overall challenges that arise when attempting to study social networks.

2.1 CHALLENGES IN STUDYING SOCIAL NETWORKS

The first challenge is the sheer size of these networks. They can involve billions of nodes and ties. Existing and traditional analysis tools often struggle to scale to these networks [31, 32]. Consequently, we need high computing power and efficient algorithms to collect, store, analyze, and process these large-scale networks [IV, V, 20, 29, 32].

The second challenge is the complex, dynamic nature of digital social networks [33]. Connections and interactions develop over time and can differ significantly in strength and type [**IV**]. Modeling these differences requires approaches beyond simple binary representations of connections [33]. The potential noise and absence of social background information about users is the third challenge that makes modeling more difficult [II, 22, 34].

The fourth challenge is developing algorithmic approaches to capture network structures and their processes while dealing with big-scale datasets. These algorithms must perform automatic labeling processes in acceptable time [**V**]. In addition, understanding the rich, multivariate metadata associated with these networks also requires novel and effective analysis and visualization approaches [30].

2.2 RESEARCH QUESTIONS

Existing methods for modeling large-scale online social networks still face several limitations. As an interdisciplinary doctoral study, this thesis integrates perspectives from computer science and linguistics to better understand social dynamics in digital environments. The key research questions addressed in this thesis are summarized as follows:

1. How can the strength of ties in large-scale, unstructured online social network data be operationalized using computational methods?

2. To what extent does network size moderate the roles of weak and strong ties in the diffusion of innovations?

3. How can user similarity be measured from user-generated social media data, and which data modalities best capture perceived similarity between users?

4. How does network size affect the accuracy of computational measures of user similarity?

5. To what extent can social media users from Nordic countries be accurately clustered into communities based solely on their network connections?

6. Do hidden, or sub-national community structures, emerge within large regional social networks beyond expected country-level divisions?

7. How can tie strength in ego networks be quantified in a robust, interpretable, and multidimensional manner using social media interaction data?

8. How can large-scale online ego networks be automatically classified along a weak–strong tie continuum using data-driven methods?

9. What structural, regional, and demographic patterns characterize ego-network clusters defined by different tie-strength profiles?

In [**I**], we revisit, from a theoretical perspective, the long-standing sociolinguistic theory of social networks [35]. We focus on the claim that weak-tie environments facilitate linguistic innovation, while strong-tie networks reinforce norms and resist change [35]. However, most prior evidence for this claim comes from small datasets derived from ethnographic observations and surveys [I,36]. Hence, it is not possible to generalize it to large social media datasets, where communication between individuals is digitally mediated, and the population frequently shifts or reorganizes connections [**I**].

Considering these challenges, in [**I**], we seek to answer whether computational methods can be utilized to operationalize network ties at scale using an unstructured dataset constructed from social media. In addition, we study whether the traditional distinction between weak and strong ties in social network theory becomes less meaningful as networks scale up [**I**]. In other words, if we consider large networks, whether the distinction between weak-tie vs. strong-tie environments undermines in promoting change or resisting it.

To address our first question in [**I**], we introduce two complementary approaches to measure tie strength in social networks and to enable sociolinguistic network analysis on a scale. We reconsider one of the proposed methods in [**I**] for measuring tie strength later in [**IV**] and [**V**], providing more detail. We adopt the method, revise and extend it. Meanwhile, in [**I**], to improve validity, we implement a bot-detection stage using metadata available in the dataset and exclude accounts with a high proportion of automatically generated content [37].

Related to the second research question in [**I**], our central finding is that network size fundamentally conditions the role of tie strength in the diffusion of innovation, particularly in small networks [**I**]. In smaller online networks, the constructed ego networks behave differently; however, as networks grow, somewhere around 120 nodes, the distinction progressively diminishes, and weak and strong ties online networks no longer show meaningful differences [**I**].

In small offline social networks, ties form for many reasons, such as friendship or collegiality [38]. However, in an online social network that is not limited by face-to-face interaction, how and why are connections established? Especially in bigger networks with hundreds of users, what is the story behind different ties? Studies show that similarity is a major factor in forming new connections on social media [39, 40].

User similarity in social media is highly subjective [II, 41]. Rather than measuring objective behavioral overlap, user similarity is often conceptualized as perceived similarity [41]. A key assumption is that perceived similarity between users mainly reflects admiration and respect rather than being measured by personal characteristics [II, 41]. In other words, users judge who is more similar to them based on their own interpretations rather than by predefined traits [**II**].

Our first research question in [**II**] is how well user-generated data can be used to investigate and measure user similarity. Also, how different data modalities and social interaction dynamics on social media can be employed to measure (perceived) similarity. Second, which data category most effectively predicts similarity? To find the answer, we consider three different categories of observable data on a social media platform: a) interaction data such as replies and retweets, b) activity history representing profile-based behavioral features, and c) linguistic content of users. We measure user similarity across these categories and compare against the ground-truth similarity dataset collected via an online survey [**II**].

We go beyond in [**II**], aligning with the investigation of the network size effect in [**I**], and ask whether the network size influences similarity accuracy. We study the correlation between network size and the similarity prediction accuracy across different categories. Our results demonstrate a negative correlation, indicating that similarity can be calculated more accurately in a smaller network [**II**].

Online social media platforms have generated large-scale interactional networks that have been increasingly studied across disciplines [42, 43]. However, these networks are often too large to be studied directly, and methods such as sub-sampling have been used to manage the volume [III, 44]. Also, most of the literature has studied national or language-specific Twitterspheres, and a regional and multi-country context is unexplored [45, 46]. The Nordic region, where countries are culturally and geographically close, provides a compelling case for exploring whether social media connectivity patterns follow geography or language [**III**]. In addition, earlier studies mostly relied on noisy approaches, such as interface languages or time zones, to identify the national Twittersphere for regional analysis [45, 46]. There is therefore a need to detect national communities more accurately, as well as to adopt a network-based approach that uses social ties to uncover the intrinsic community structure in a large regional Twitter network.

In [**III**], we seek to answer how accurately Twitter users can be clustered based solely on their social connections. We investigate whether clustering results based on Twitter friends' relationships align with users' home countries in the Nordic region. By asking this question, we aim to address a broader issue: to what extent online connectivity mirrors national identities, even in geographically close and culturally similar countries in the Nordic region.

Our second research question in [**III**] focuses on hidden and sub-national clusters. We examine whether additional, previously unknown clusters exist beyond the expected country-based divisions (5 clusters representing the five Nordic countries). This also includes whether users from one country or a country cluster can

be further subdivided into two or more stable clusters based on interaction ties representing friendship in a social network.

In [**III**], we finally focus on the suitability of the clustering criteria and evaluate our methodological robustness. By using different cost functions for clustering, we seek to determine which one best captures community structure in the Nordic Twittersphere [III, 47]. Particularly, we compare three cost functions and evaluate their performance [III, 47].

In the literature on computer-mediated communication (CMC), there is increasing interest in large-scale social media data; however, these datasets are mainly treated as text corpora, with their fundamental network nature being ignored [IV, 48]. In addition, existing studies on measuring tie strength in online networks typically rely on a single indicator, such as shared friends, thereby oversimplifying the multidimensional nature of social tie strength and failing to capture many of the factors that influence it [20, 29, 35]. Also, existing approaches treat tie strength as a pairwise concept, making it difficult to rank and compare networks by tie strength [**IV**]. There is, therefore, a methodological gap for a robust, interpretable, and scalable computational approach to quantifying tie strength in a complete social structure, such as ego networks, that we aim to fill in [**IV**] through modifying the basic model we introduced in [**I**].

In [**IV**], the first question we ask is: to what extent can the tie strength of ego networks constructed from computer-mediated communication be quantified using interaction metadata extracted from social media? The question targets the lack of systematic and interpretable measures for describing how weak or strong an individual network is. In the second research question in [**IV**], we concentrate on the multidimensionality of tie strength. In particular, we contrast one-dimensional approaches and explore how measures derived from interaction patterns and network topology jointly influence tie strength in networks.

Addressing previous research questions in [**IV**] allows us to investigate how large-scale compute-mediated communication data may shape language use and social behavior. After developing an algorithmic approach that captures the multidimensional nature of tie strength and places individual social networks along a continuum from weak to strong tie structures, we can then ask what new types of linguistic and social analysis become possible. By incorporating structurally informed measures such as tie strength, we treat social media not merely as a text corpus, opening new avenues for examining language variation, diffusion, and social behavior at scales that were previously inaccessible.

We know that tie strength is a central concept in social network theory, influencing information diffusion, innovation, and social support [23, 49, 50]. Yet it remains difficult to model in large-scale datasets constructed from online networks [**V**]. In [**V**], instead of relying on thresholds or domain-specific heuristics, we introduce a multidimensional approach that uses the revised measure set proposed in [**IV**] to label each network by tie strength. In [**IV**], we focused on measuring the strength of the entire ego network rather than classifying individual ties; however, what remains missing is an automated, scalable labelling method for large-scale social network datasets. In [**V**], we address this gap by proposing a platform-independent approach and bridging social network theory and data-driven clustering. The proposed framework in [**V**] allows labelling digital ego networks based on measurable structural and instructional properties.

In [**V**], we seek to answer how online ego networks can be automatically clustered. Doing so, we aim to move beyond the binary distinction between weak and

strong ties toward a graded categorization of networks along a weak-strong tie continuum. This question seeks to address the lack of a generalized platform-agnostic method for labeling networks from online social media. The second research question in [**V**] concerns the structural patterns that may emerge across clustered tie-strength categories. We concentrate on statistical and social characteristics distinguish the clustering results. We examine whether there are systematic regional and demographic differences among the detected tie clusters.

To address the existing gaps mentioned in this section, a social media platform is required that provides large-scale, real-world interaction data. Among available platforms, Twitter (used to) offers a unique environment for studying social interactions, information flow, and language use. Its open data policy, rich interaction patterns, and global user base make it suitable for testing social network theories and developing computational models at scale.

3 Twitter as a research laboratory

Twitter, currently known as X, is the primary data source of this thesis and its published papers. Twitter is a social network created in March 2006 and allows its users to share text and URLs (up to 280 characters), images, and videos. Every account (user profile) on Twitter has a list of *followers*, those who are following the account, and a list of *following* (friends), those the account follows. Accounts on Twitter could be either public or private. For the public accounts, their content would be visible to others, while for the private accounts only their follower can see their contents.

Twitter previously maintained an open policy regarding data accessibility. The platform used to offer various free Application Programming Interfaces (APIs), which enabled researchers to collect large-scale data from the Twitter archive. This converted Twitter into a valuable research laboratory and an ideal environment for studying large-scale social media data and examining societal phenomena in a digital context.

Following the Cambridge Analytica controversy in 2018, an era often referred to as the "APIcalypse" began, during which many social media platforms significantly restricted data access [51, 52]. Twitter was one of the last big companies to follow suit. It maintained free API access for years, only ending it after its rebranding in mid-2023, when it introduced a paid model instead. For the purposes of this thesis, we utilized Twitter's now-discontinued Academic API, which once allowed researchers to gather historical data reaching back to the platform's early days.

Working with large-scale social media data inevitably raises questions about ethics, privacy, and data ownership. Public posts may seem open to analysis, yet the boundaries of what is acceptable to collect and interpret are not always clear. In the European Union, the *Digital Single Market Directive* (2019/790)[1] has made text and data mining for scientific research purposes legally possible, offering a more transparent framework for responsible data use. This thesis follows the ethical responsibility principles by focusing only on publicly available information and anonymizing it so that it cannot be traced back to individual users.

3.1 EGO NETWORK

A key concept we focus on is the *ego network* (see Figure 3.1). Ego networks are the foundation and primary characteristic of individual networks and are essential when analyzing human behavior on social networks [53,54]. As illustrated in Figure 3.1, an ego network is a social structure centered around an individual (*ego*), its direct connections (*alters*), and the connections between alters.

Ego networks provide the analytical granularity necessary for understanding tie strength and user similarity, which is explored throughout this thesis. Studying them offers a close look at individual behavior while still revealing broader patterns of connection. In addition, ego networks are the personal layer of a social network,

[1] Directive (EU) 2019/790 of the European Parliament. Retrieved 7 November 2025, from https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng

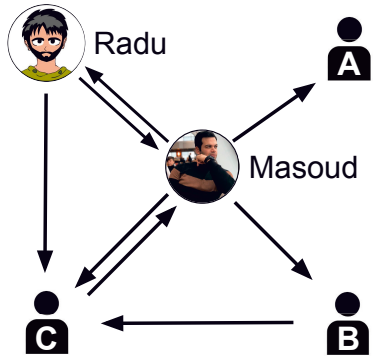


Figure 3.1: Masoud's ego network consists of 5 nodes and 8 directed edges.

offering not only who someone is connected to but also how those connections interact with one another.

In Figure 3.1, Masoud is the ego node and follows Radu, A, B, and C. While A, B, C, and Radu represent the alters, with only Radu and C following Masoud back. The arrows shown in Figure 3.1 capture these one-way or mutual relationships, indicating that a directed edge points from the user who follows to the one being followed. From Masoud's point of view, then, Radu, A, B, and C are his *friends*, whereas only Radu and C count as his *followers*.

3.2 LOCATION ON TWITTER

Accessing location is a long-standing and important challenge while working with online data. For instance, Tabarcea et al. [55] provided a good summary of how location is used in search engines, where the key is first to extract and access the location. On Twitter, users can share their geographical location through several channels, such as tweet text, profile bio, or geo-location-enabled tweets. When analyzing Twitter data considering the geographical location, two main questions usually arise. The first is how easily a user's location can be inferred, and the second is how trustworthy each method is for doing so. The following walk-throughs the three different channels that can be utilized to identify location and outline what each one offers, along with its shortcomings.

The first channel that enables accessing geographical locations is the tweet text. In a study by Tabarcea et al. [56], the authors utilized street-name prefix trees to extract location from free-form text. However, on Twitter, the location is rarely exact down to the street address but is more likely to be a casual mention of the place (see Figure 3.2). On Twitter, as shown in Figure 3.2, users may include location names directly in the text of a tweet. Extracting that information, however, is not straightforward. It first requires *natural language processing* (NLP) methods to detect potential place names and, second, to validate them. The following explains some of the challenges that might arise while extracting the location for their tweet-text

The first challenge is that many locations share the same name, which can make interpretation problematic. For instance, in the tweet shown in Figure 3.2, the word *Paris* might refer to Paris, France, or just as easily to Paris, Texas. Without additional context, there is no reliable way to distinguish which one the user meant and actu-

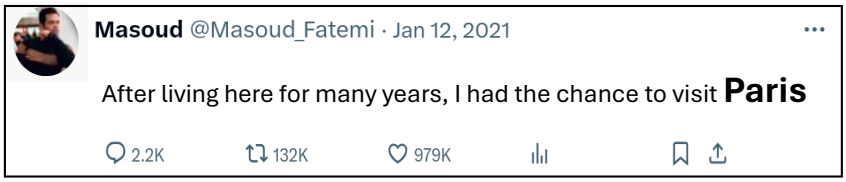


Figure 3.2: A tweet including a location pointer.

ally visited. In this scenario, the final result is a precision vs. recall decision-making problem, as explored by Fränti et al. [57], but in a different domain. A greedy selection of names leads to high coverage but many incorrect locations (high recall but low precision), whereas too strict verification results in only a few, but more certain, locations (low recall but high precision).

Another challenge in extracting location from tweet text is the presence of common words that can also represent place names. For example, *Orange* could refer to a color, a fruit, or a city in southern France. In addition to this ambiguity, users often employ abbreviations or nicknames, such as *Philly* for *Philadelphia*, which further complicates the task. In such cases, contextual or localized knowledge is necessary to infer the actual location. Finally, even when a location name is successfully detected in the text, its mention does not necessarily imply that the author resides there. For instance, in Figure 3.2, the user states he visited Paris, but there is no indication that he lives there or was in Paris at the time of posting the tweet.

Considering these challenges, extracting location from the tweet text was not applicable and was out of the scope of this thesis, and we did not consider it during this research journey.

The second channel that can be utilized is extracting the location from the Twitter profile. In every Twitter profile, as demonstrated in Figure 3.3, there is a location field that allows users to share their location with others. However, this location field is not reliable. It is a free text field that allows users to enter locations in different formats and mention locations that do not exist or are fictional.

The third channel is utilizing geo-tagged tweets. On Twitter and under the user profile settings, there is a feature that, if the user activates it, a geolocation tag will be added to every single tweet that the user posts.

Figure 3.3: A Twitter user profile.

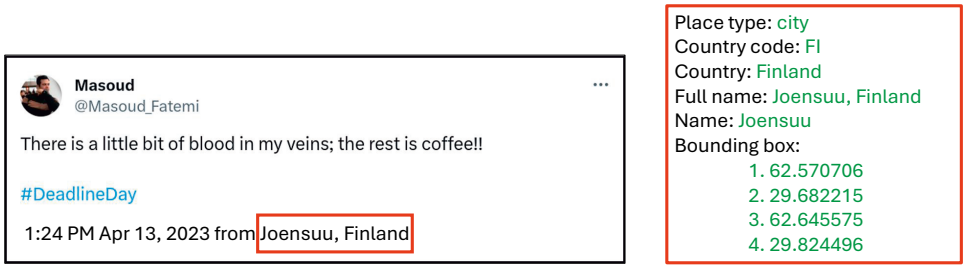


Figure 3.4: A geo-location-enabled tweet and how the location information looks like on Twitter (left) and how it looks like in the collected data using the API (right).

Fazal et al. [58] explored a web crawler to retrieve geo-tagged pages, and the resulting count was very low, somewhere <0.1%. Social media, Twitter in our case, however, is dynamic, and content is more often created on the go. Location is easier and more natural to attach to Tweets than to static web pages, which are rarely visited when people are visiting the places. In fact, social media was considered a significantly more promising source for location-tagged content for content creation than the web [59].

Even though due to privacy preservation the majority of Twitter users do not use this feature [22], but still, some of them share their tweet locations. We need a sufficient sample size from the population to make reliable conclusions about our study subjects; in contrast, location-based applications should have high coverage of relevant locations [59].

Figure 3.4 (left) illustrates a geolocation-tag-enabled tweet and how its location information is displayed on the Twitter platform. Figure 3.4 (right) displays the same tweet's location data as extracted for our data repository collected via the Twitter API. In contrast to the location information from a user's profile, Twitter provides geo-tagged tweet location information in a standardized format, including the country ISO code[2] and geographical coordinates represented as a bounding box defined by the minimum and maximum latitude and longitude values.

[2]List of country codes by alpha-2, alpha-3 code (ISO 3166). (n.d.). Retrieved 28 October 2025, from https://www.iban.com/country-codes

4 Methods for modelling online social networks

In this section, we briefly review different approaches toward modelling online social networks. These models have evolved from manual, theory-driven measurements in small offline communities to computational, data-driven approaches that can handle large-scale digital environments. Early work on measures of tie strength focused on individual ties in small networks, while more recent approaches, such as our proposed method, enable quantifying entire ego networks.

4.1 MANUAL MEASUREMENT

The traditional sociological and sociolinguistic approaches fall in this category. The core characteristics of these approaches are that they mainly rely on human judgment, surveys, interviews, and ethnographic observation [15, 60–62]. In these approaches, the strength of the tie is inferred from the time people spend together, the emotional intensity of the relationship, intimacy, and reciprocal support [15, 60, 61]. Considering the methods that been used for data collection in this category, such as ethnographic observations, the networks are typically small, somewhere around 30–50 nodes, which are deeply contextualized and domain-dependent [60, 63].

In these traditional methods, researchers typically used a name-generator survey to detect social ties, in which participants listed people they regularly interacted with [60,64,65]. Interactional data among individuals are sometimes manually coded by the researchers based on the observed or reported events. Tie strength is also typically assessed using self-reported evaluations based on measures such as interaction frequency or relationship duration. Tie strength is commonly classified into binary or ordinal categories, such as weak ties vs. strong ties [15, 60, 65].

These methods have both strengths and limitations. On one hand, these methods emphasize interpretability and high contextual validity. Since they are based on observations and self-assessment, they capture the social meaning behind the relationships. Also, like Granovetter [15] and Milroy's research [60], they have a strong grounding in social theory. On the other hand, these methods suffer from poor scalability to large populations and are mainly limited to small networks. They are also costly in terms of the required time for data collection and highly subjective. Moreover, due to their nature for data collection and data encoding, they are difficult to replicate across different contexts or domains.

4.2 COMPUTATIONAL ESTIMATION

Manual measurement and labelling failed to scale up, especially with the emergence of online platforms generating massive interaction data [I, IV, V, 66]. Accordingly, there was a need for algorithmic and data-driven approaches capable of analysing large, rich datasets [I, 67]. These methods aim to infer tie strength indirectly from observable behaviour. For instance, they estimated tie strength for individual ties using proxies, such as interaction frequency [29] or share of common friends [68].

The majority of methods that use proxies to estimate tie strength rely on single-dimensional metrics [20,29,68]. These approaches typically operationalize tie strength using measures of interaction frequency or interaction intensity. Interaction frequency may refer, for example, to the number of phone calls, text messages, or direct messages exchanged between two users [18, 23, 29, 69]. Interaction intensity can also incorporate structural features of the network, such as neighbourhood overlap or the proportion of mutual connections [20,21,68]. These methods typically use predetermined thresholds for classification, such as median or heuristic cut-offs, to label ties as weak or strong. In addition, these models focus on edges rather than entire social structures [18,23].

Unidimensional estimation approaches overcome the drawbacks of manual labelling methods. They can scale up to handle large datasets with many users. However, they consider tie strength unidimensional, overlooking its complex nature. Also, they rely heavily on ad hoc thresholds, rendering their results context-dependent and impossible to generalize. Finally, they are edge-level methods and cannot produce a single interpretable value for an ego network.

4.3 A MULTIDIMENTIONAL APPROACH

Here, we briefly elaborate on our multidimensional approach, which is a platform-agnostic network-level method for modelling online social networks using a predefined measure set [**IV**]. The initial version of this model relied on six measures and was applied in [**I**]. Later, in [**IV**], we improved the approach, expanded its measure set to eight, and studied it in more detail. Through statistical analysis and measure validation, we identified the most effective measures proposed in [**IV**] and employed them in [**V**] to cluster tie strength in online social networks.

Our approach builds upon this idea that social relationships are inherently multifaceted and have many sides; therefore, they cannot be reduced to a single metric [I,IV,V,35]. Tie strength should reflect the complex nature of a network dynamic and capture many aspects, such as the instructional, structural, and positional properties of nodes and edges [**I**, **IV**, **V**]. Our proposed approach moves from pairwise tie strength to whole-ego network strength, from single indicators to multidimensional indices, and from binary classification to a continuous strength scale. We assume that network strength originates from the combined structure of interactions, not from isolated edges, and that ego networks can be meaningfully placed on a continuum from weak-tie to strong-tie networks [**IV**].

We introduce the *network strength index* or NSI, a single interpretable score that maps each ego network to a spectrum of weak-to-strong-tie networks [**IV**]. We calculate NSI based on eight measures to quantify tie strength at the network level. These eight measures are as follows:

1. *Interaction strength* (IS): A weighted value based on the frequency of interactions between nodes in the networks. Higher IS values indicate a stronger tie network.

2. *Relative interaction strength* (RIS): The relative version of the IS measure. It calculates the proportion of interactions between alters relative to those between alters and egos. RIS interpretation aligns with IS; higher values indicate a stronger tie network.

3. *Social similarity* (SS): The number of common friends between pairs of nodes in the ego network. As with SS, the higher the value, the stronger the tie in the network.

4. *Outliers* (OUT): The proportion of alters that become isolated when the ego node is removed. For OUT, lower values happen in stronger tie networks.

5. *Asymmetry closeness centrality* (ACC): The absolute difference between incoming closeness centralities and the outgoing closeness centralities. A lower value for ACC indicates a stronger tie network.

6. *Average of betweenness centrality* (ABC): The average betweenness centrality for all the nodes in the network. Lower values of ABC mean a stronger tie network.

7. *Range of betweenness centrality* (RBC): The difference between the maximum value and the minimum value of betweenness centrality of nodes in the network. Similar to ABC, lower RBC values also indicate a stronger tie network.

8. *Density* (DENS): The number of available links in the network compared to the total possible number of edges. Higher values for DENS mean a stronger tie network.

Among these eight measures, IS and RIS are based on interactions between nodes (altars and ego) within the network. The SS measure is derived from set theory and is based on Jaccard similarity [70]. ACC, ABC, and RBC are centrality-based measures from the field of graph theory. OUT and DENS are topological and structural measures we extract from each ego network.

After extracting these measures per network, reverse scaling is needed for OUT, ACC, ABC, and RBC to ensure consistency, so that in all eight measures, higher values denote a stronger tie network. Then we normalize the measures to ensure they have the same impact on the final value. Finally, to calculate the NSI, we average these eight measures per network, yielding a value between 0 and 1, where 0 is the weakest and 1 is the strongest.

Our method matters because it avoids relying on a single heuristic or threshold for tie-strength classification. It allows us to assign a single value per network that describes the tie strength of the entire social structure. By doing so, we can compare and rank ego networks based on their tie strengths. Finally, the whole pipeline is explicitly designed to handle large-scale computer-mediated communication data, specifically constructed from social media.

5 Network modelling and analysing behaviour

In this section, we first briefly explain clustering and graph clustering, which have been used as tools throughout this thesis. Next, we shortly review the five articles on which this thesis is based on. We go through papers one by one and explain each paper's main focus, motivation, research questions, methods, and key results.

5.1 CLUSTERING

Clustering is an unsupervised machine learning method, meaning it operates on data that does not include predefined labels [71]. In clustering, the goal is to group objects, datapoints, entities, so that similar items are placed in the same cluster while dissimilar items are placed in different clusters [72]. In other words, a cluster is where objects inside the cluster are similar to each other compared to those outside the cluster (in other clusters) [73].

There are massive, complex datasets that are nearly impossible for humans to manually analyse and extract information from [47]. By grouping related items, clustering facilities simplify these large datasets [47]. By doing so, clustering enables the identification of hidden patterns, such as multimorbidity groups, in healthcare data [73,74].

For a given dataset as $X = \{x_1, x_2, \ldots, x_N\}$, where N is the number of datapoints, clustering aims to find a partition as $P = \{P_1, P_2, \ldots, P_k\}$ of X into k disjoint clusters, and the centre of each partition as $C = \{c_1, c_2, \ldots, c_k\}$ [75]. Figure 5.1 indicates a

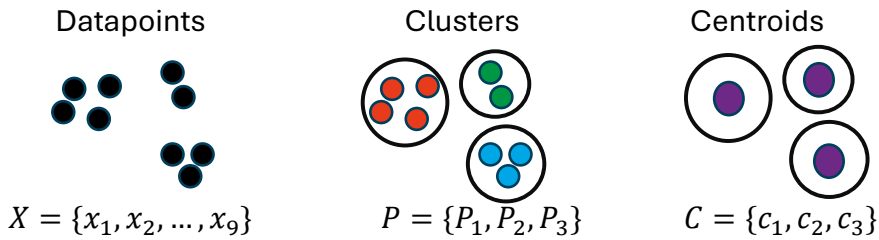


Figure 5.1: A dataset (left), clustering results (middle), and the centroids of clusters (right).

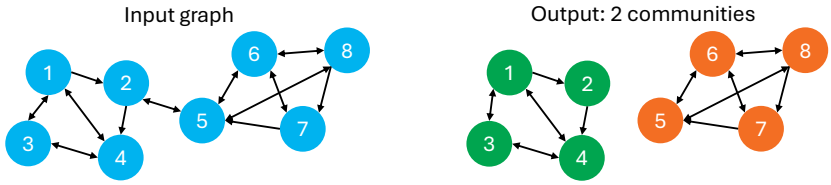


Figure 5.2: An input graph including 8 nodes (left) and the graph clustering results into 2 clusters/communities (right).

K-means is a well-studied partition-based clustering algorithm [75] that is used in [**V**] to cluster Twitter ego networks into different tie-strength categories.

Graph clustering, or community detection, as illustrated in Figure 5.2, is the partitioning of a network's nodes into clusters, where nodes within each cluster are more connected to each other than to nodes outside the cluster [47]. In other words, there are more connections within the nodes inside each cluster than outside it [47]. Figure 5.2 represents an input graph including 8 nodes, which is clustered into two clusters/communities. Graph clustering can be used to simplify large networks, such as disease co-occurrence networks, so humans can analyse them and the information from them [47,73].

M-algorithm is a graph clustering method directly derived from k-means [47]. There are three main differences between standard k-means and the M-algorithm. First, the standard k-means works on numerical data, meaning it requires numerical input values to calculate centroids (cluster means). In contrast, the M-algorithm cannot directly calculate the mean from the input data, since the input is a graph (nodes and links). Second, in k-means, distance measures such as Euclidean distance are used to evaluate how to move a datapoint between clusters. However, in the M-algorithm, a step called the delta approach is used to evaluate moving nodes between clusters and their impact on the cost function. Third, a common weakness of standard k-means is falling into local optima; however, to avoid this problem, M-algorithms randomly merge two clusters and split one cluster before fine-tuning the result [47].

We utilized graph clustering, particularly M-algorithms, in [**III**] to detect country clusters in a network dataset compiled from the Nordic region.

5.2 LANGUAGE CHANGE

Language change is the universal and constant evolution of the language system [76]. Birth and adoption of new words (*neologisms*), spelling, and grammatical structures are all examples of language change [**I**]. For a change to succeed, speakers (social media users in our case) must first come into contact with it and second decide to use it [I,77]. *Innovation diffusion* is the social process that spreads language change from its origin (innovators) to a broader community, where they decide whether to adopt it and whether it becomes established as a norm [**I**]. That is why a successful language change is considered a social phenomenon.

One major limitation in the literature on weak-tie hypotheses in sociolinguistics is that these studies have historically relied on small-scale ethnographic data, often collected through observations and interviews [**I**]. In [**I**] Our primary moti-

vation was to address this gap by modernizing the weak-tie hypothesis through analysis of large-scale online social networks. In particular, we tested whether the traditional and long-standing distinction between weak-tie environments (promoting language change) and strong-tie environments (resisting it) disappears when we consider large networks of mobile individuals [**I**].

To achieve our goal, as the first step, we needed to collect and form a dataset of large online networks. We manually handpicked 10 non-academic Twitter accounts that primarily post content in English from the metropolitan areas of Helsinki and Stockholm [**I**]. We used the Academic Twitter API to collect their entire networks as well as all recent messages posted by these accounts, up to the last 3,200 (the API's limit). We also applied more filters, such as having more than 300 friends and followers combined on the platform.

As of the second step, we needed to develop a computational method to estimate tie strength at scale. The proposed method had two main features. First, it relies on the idea that, in contrast to existing models, tie strength is a multidimensional concept influenced by many factors [**I**]. Second, it should be able to handle large networks collected from social media and vast datasets. Accordingly, we borrowed concepts from graph theory and set theory, and via analysing the collected interactional data, we extracted six measures from each network [**I**]. After transforming all the extracted measures to the same scale and normalizing them, we calculated the average values for each network as the tie strength value. We sorted the networks by the calculated values and labelled the top 5 networks with the highest values as strong-tie networks, and the rest as weak-tie networks.

For the third step, we needed to define innovation and encode it in a way so that we can track and analyse it within the networks. Since we are dealing with social media data, mainly text, we focused on defining and extracting innovation from text content [**I**]. In more detail, we used a combination of linguistic measures, such as the use of contracted forms (e.g., won't, I'm), and grammatical changes that are mainly used by non-native speakers [**I**]. We considered these digital measures as modern markers of how new language habits disseminate.

In the last step, we applied statistical testing to compare the identified weak and strong-tie networks against the encoded and extracted linguistic measures. Our results demonstrated that the traditional distinction between weak and strong-tie networks for promoting or resisting change disappears in online networks with more than 120 participants [**I**].

5.3 USER SIMILARITIES

In social network analysis, identifying users who are more similar to each other based on specific traits or sharing behavioral characteristics is called user similarity [II, 78]. However, user similarity is highly subjective and, in most cases, rather than having common characteristics, it involves admiration or respect between users [II, 79]. There are also different versions of user similarities, such as self-view, perceived, and peer-view similarities [II, 80].

User similarity matters in social network analysis, as it facilitates understanding of the social structure and helps predict user behavior [II, 81]. Common characteristics between users are often what initiate building a connection in the first place [**II**]. Also, users who are more similar to each other are more likely to share prefer-

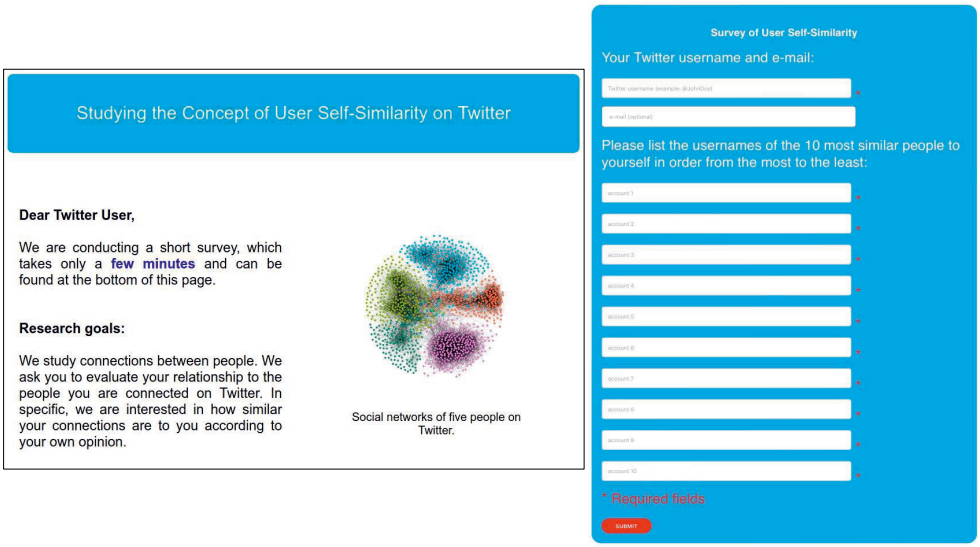


Figure 5.3: The online platform we designed for collecting ground-truth data from Twitter users. We published the survey through university channels. The collected answers and data then been anonymized and stored on the School of Computing servers at the University of Eastern Finland.

ences, enabling social media owners to improve their recommendation systems and personalized suggestions [82].

In [**II**] we focused on identifying what makes Twitter users feel similar to each other. In other words, we concentrated on perceived similarity. Previous studies argued that it is a highly subjective phenomenon [**II**]. Consequently, instead of focusing on a specific trait to measure perceived similarity as in the standard definition, earlier studies mainly focused on measuring the extent to which users admire others [**II**]. To study this phenomenon in [**II**], we further explored the similarity component. We considered both user-generated interaction data and linguistic features, as well as activity history, to measure perceived similarity. In particular, we first asked to what extent user-generated data on Twitter can be used to investigate user perceived similarity. Second, what user-generated data most effectively predicts perceived similarity?

To answer these questions, we first needed ground-truth data. We designed an online platform[1] (similarity survey) and collected ground-truth data directly from Twitter users. We asked survey participants to rank the top 10 users most similar to themselves. 14 Twitter users participated and filled in the survey. For those who participated, we collected their entire network, as well as the latest 3,200 messages/tweets for each user in their network. Figure 5.3 presents information from the similarity survey (left) and the similarity form we asked participants to fill in and submit (right).

Second, we developed three quantitative models to measure the similarity of Twitter users using data collected from their accounts. We asked Twitter users to provide a list of accounts similar to themselves and then compared this list against

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

three models we developed to measure similarity. We measured similarity a) using the interactions between an ego node and its alters, b) analyzing the hashtag set for each ego node and its alters, c) profiling activity history for each ego node and its alter and then measuring the distance between profiles [**II**].

Our results in [**II**] suggested that using the interactions that an account has with its alters is the best way to measure similarity between Twitter users. In other words, measuring user similarity based on interactions between users outperforms both hashtag and activity history similarity. Our results also demonstrated that the accuracy of measuring similarities based on interactions, hashtags, and activity history is negatively correlated with network size (number of nodes). Finally, a gender analysis showed that, for male networks (male ego nodes), we can measure similarity more accurately than for female networks (female ego nodes), regardless of the method used [**II**].

5.4 DETECTING COMMUNITIES

The literature studied Twitter data from different countries, mostly focusing on a single country or a specific language [III, 44]. In [**III**], as an unexplored context and to address the limitations of national Twitter studies, we focused on multiple countries. By taking a regional and multi-country perspective, we concentrated on the Nordic region. The Nordic countries form a theoretically interesting case as they are geographically close, culturally similar, and partially linguistically related, with the exception of Finnish, yet different countries [**III**].

In [**III**], we analysed whether we could cluster geographically labelled Twitter user networks and to what extent the clustering results align with users' home countries in the Nordic region. We also tried to extract and explore whether there are any hidden or additional cluster(s) beyond the five-country clusters in the data. Also, utilizing three different objective functions, which one is the most suitable for clustering a geographically sparse social network dataset? Finally, we analysed whether there is content similarity between clusters from different countries in the region.

To meet our goals, the first step was data collection and constructing a geographically labelled network dataset. In this regard, we used the Nordic Tweet Stream[2] or NTS [83].

NTS is a web application developed at the University of Eastern Finland as part of a project funded by the Research Council of Finland and its Research Infrastructure program. It allows users to search, subset, visualize, and download user-generated social media data from the Nordic region. The NTS dataset contains nearly 74 million messages from over 888 thousand user accounts from January 2013 to May 2023. Figure 5.4 represents the NTS homepage. The author is a member of the NTS development team, and the tool is hosted by the IT Center for Science (CSC) in Finland.

We extracted all users whose tweets were included in NTS and collected their information from Twitter. In the second step, we filtered the collected users to verify their locations, excluded suspicious users with highly abnormal activity patterns, and formed a directed network for the remaining accounts.

In the third step, we applied the M-algorithm (see section 5.3) to cluster our network dataset into 5 clusters, corresponding to the five countries in the region [III, 47]. We utilized three cost functions to compare how well each objective function

[2]The NTS web application is available at: https://nordictweetstream.fi

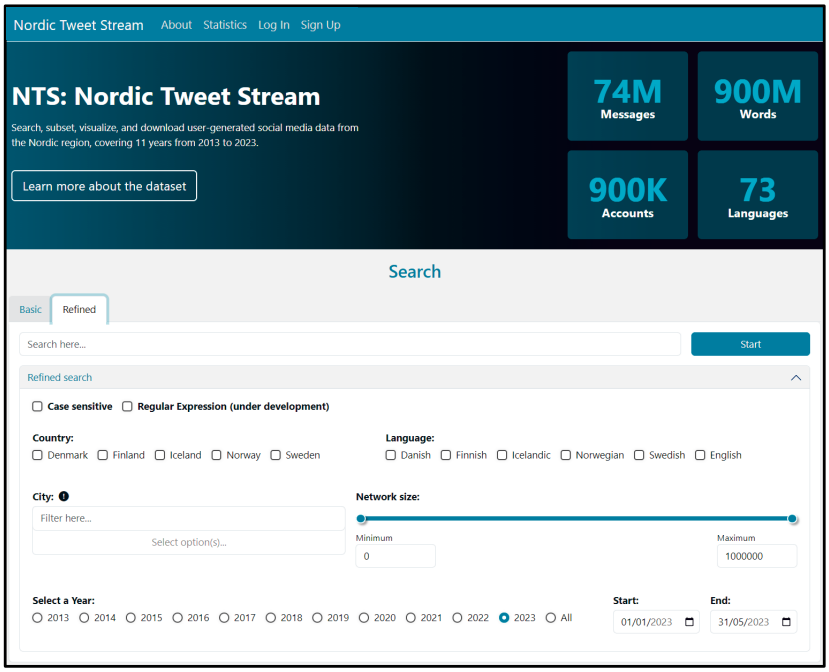


Figure 5.4: NTS web application home page.

performs, and its results match the ground-truth labels of the user's county [III, 47]. Our results demonstrated that conductance outperforms the other objective functions, and its results best align with the users' home country clusters [III, 47].

In the fourth step, to analyze potential hidden clusters in the data, we tried to place the data into 6 clusters [**III**]. Figure 5.5 illustrates the ground-truth network at the top and the clustering results from adding an addiction cluster across 6 different runs. As shown in Figure 5.5. Adding the 6th clusters made the results very unstable, and the location of the 6th clusters changed in different runs of the algorithm [**III**]. The instability in adding an extra cluster demonstrated that there are no 6th clusters naturally.

We also examined the existence of sub-country clusters. However, the outcome suggested that there is no meaningful sub-country clustering in the data based on geographic location [**III**].

In the last step, we analysed the content of clusters detected by the M-algorithm. We analysed hashtag statistics for each country cluster, as well as the most used hashtags for each country, to identify the dominant theme [**III**]. In addition, we calculated overlaps among the hashtags used in each cluster pair to measure content similarity. We aimed to analyse how closely content similarity follows the connection patterns in our network dataset. Our results demonstrated that countries that are strongly connected do not necessarily share a high number of hashtags or use the same set of hashtags [**III**].

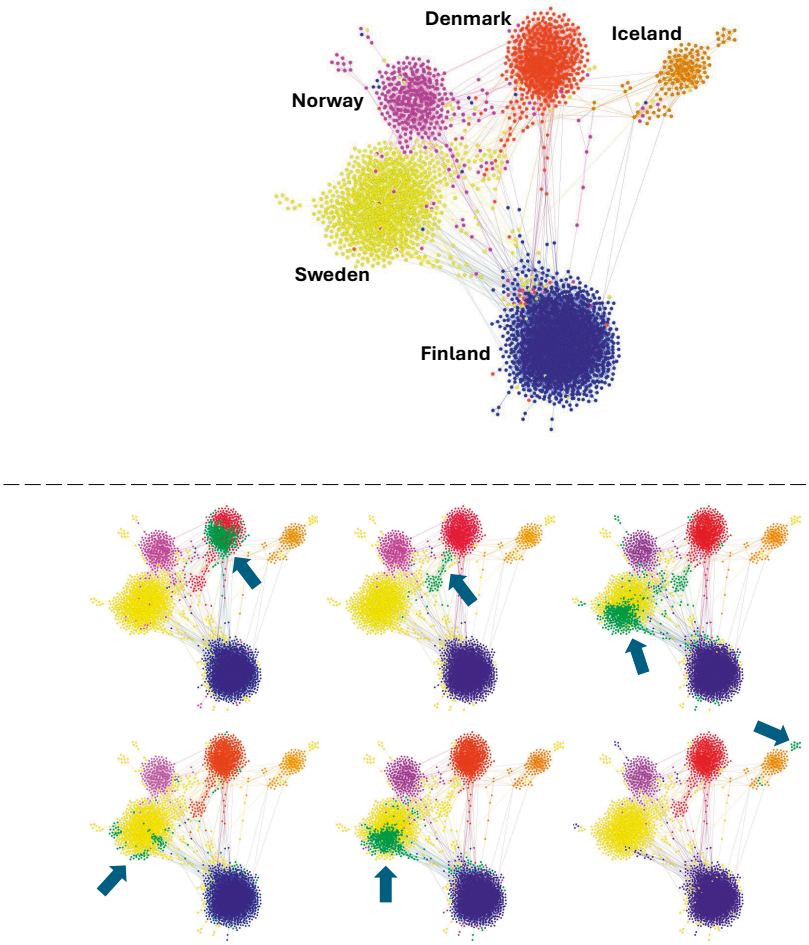


Figure 5.5: The ground-truth Twitter network from the Nordic region (top). Clustering results from 6 different runs of M-algorithms and clustering data into 6 clusters (bottom).

5.5 MEASURING TIE STRENGTH

In **[IV]**, we proposed a computational approach to construct an ego network from computer-mediated communication (CMC) data. We utilized a massive geographical network dataset, the digital social network corpora (DSN corpora), collected from Australia (AU), the United Kingdom (UK), the United States (US), and the Nordic Twitterspheres **[IV]**. Figure 5.6 (top) illustrates different parts of the DSN corpora. The dataset includes three native English-speaking countries and the Nordic countries. Figure 5.6 (Middle) shows the geographical distribution of ego nodes in DSN corpora from Australia, the United Kingdom, and the United States. Finally, Figure 5.6 (bottom) presents ego networks and their interconnections from three cities in DSN Britain, drawn from the DSN corpora. For each city network, each color represents an ego network.

We introduced the network strength index (NSI), which enables quantifying

tie strength not for a single tie or node in the network but for the whole social structure, such as an ego network [**IV**]. NSI is a revised and extended score compared to the method utilized in [**II**]. What makes NSI more valuable is that it is platform-independent and can be applied not only to Twitter networks but also to any platform that supports the formation of directed social networks, such as Bluesky [**IV**]. We also demonstrated how enriched network information, by adding network strength values, can be used to study linguistic behavior and illustrated this through a case study on swearing on Twitter.

In computer-mediated communication research, social media is often considered only a textual corpus and overlooks the networked nature of data [**IV**]. Also, a robust, standardized approach to forming social ego networks from interactional data was missing. In addition, as explored in [**II**], the existing literature treated tie strength as a unidimensional concept applicable to a single node or tie, whereas we argued the opposite and introduced an approach that allows assigning strength values to a complete social structure and ranking them [**II**, **IV**].

In [**IV**], we mainly focused on the possibility of forming ego networks from interactional data collected on Twitter and on the extent to which we can systematically measure the tie strength of these structures. In more detail, we aimed to measure tie strength for each ego network utilizing eight measures and to place each network on a spectrum of weak to strong ties [**IV**].

We collected and created a large-scale network dataset. We streamed geolocated tweets from AU, UK, and the US and extracted the users who published them. For the Nordic region, as we did in [**III**], we used the users whose tweets were included in the NTS tool. After a few filtering phases aimed at identifying *genuine* human accounts, we collected ego networks, interactional data, and messages for accounts that passed the filtering [**IV**].

In [**II**], we utilized 6 measures to calculate network tie strength. In [**IV**], after forming the dataset, we defined a measure set, a) comprising the 6 measures used in [**II**] and revised one of them, b) one new proposed measure, and c) one taken from a graph theory concept. In total, we extracted 8 measures per network and, after processing to ensure consistent interpretation, calculated the NSI. This single value represents the tie strength per network and ranges from 0 (weakest) to 1 (strongest) [**IV**].

To validate our measure set, we generated random ego networks as a baseline to compare the distribution of our measures across real-world and randomly generated ego networks [**IV**]. In addition, we examined the correlations between the 8 measures and network sizes and degrees to assess the robustness of these measures. Moreover, as a case study, we employed the NSI to test how swearing, as a linguistic feature, varies with tie strength and network size [**IV**].

Our results demonstrated that using the proposed method and NSI value, we can produce a single interpretable value for each ego network, enabling us to compare, rank, and group networks across large datasets by their tie strength [**IV**]. The validation stage indicated that not all the measures used are equally informative, with some being best at distinguishing between random and real-world networks. While some of the measures, especially those more based on user interactions, are more robust across networks of different sizes and degrees, and better support comparability [**IV**].

In addition, comparing real-world networks with randomly generated networks revealed that the former tend to have more weak-ties, with many networks organized around the ego and fewer alter-alter connections [**IV**]. Finally, the linguistic

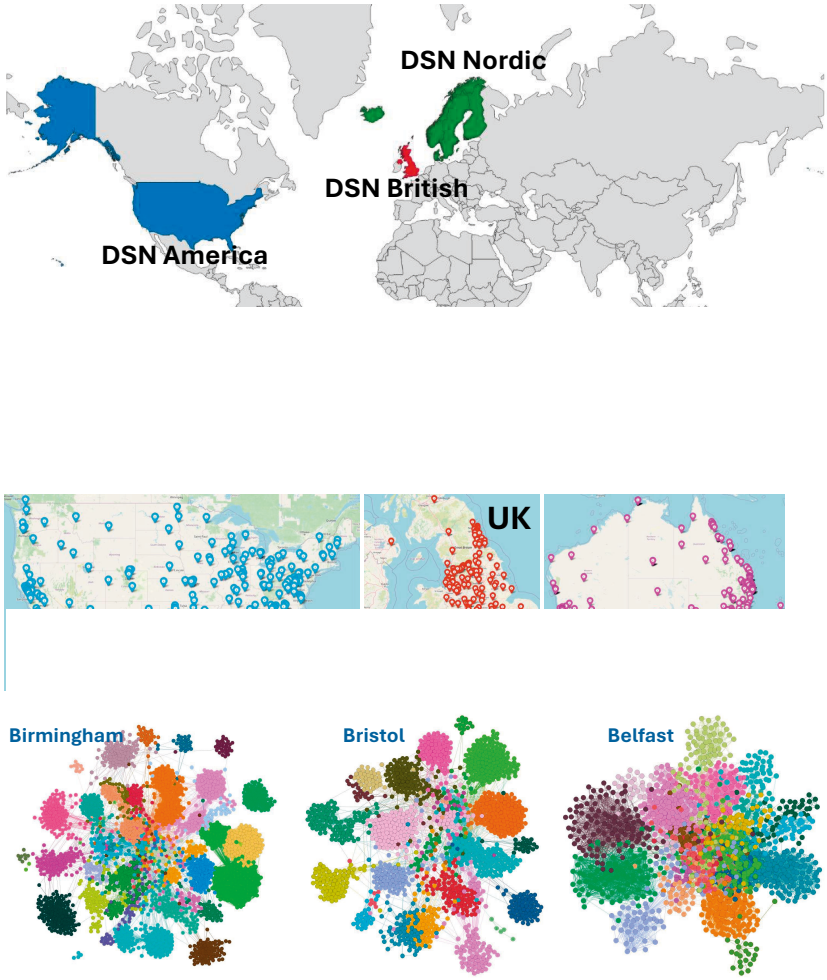


Figure 5.6: DSN corpora structure. Top, different regions and parts of the DSN corpora. Middle, The geographical distribution of ego nodes in AU, UK, and US. Bottom, ego networks from three cities in DSN Britain.

case study indicated that swearing with F-words varies with network properties, with a higher probability of using such items in weak-tie networks [**IV**]. Also, in the case of F-word usage, size matters: very small networks show a lower usage rate [**IV**].

5.6 CLUSTERING TIE STRENGTH

[**V**] is the last paper included in this thesis. We received the best paper award for **V** in the ISKE 2025[3] conference in November 2025 in Shunde, China.

In [**V**], we proposed a scalable, automated method for classifying ego networks based on tie strength. In [**V**], using the four best measures calculated in [**IV**], we

[3]International conference on intelligence systems and knowledge engineering (ISKE): https://iske2025.com

treated each ego network as a datapoint in a four-dimensional space. Then, we applied clustering and placed ego networks into four groups: weak, moderately-weak, moderately-strong, and strong. We aimed to move beyond traditional threshold-based approaches and manual labelling to categorize networks as weak or strong.

One major limitation of methods for detecting and labelling networks by tie strength is that they rely on ad hoc thresholds or manual labelling. In addition, prior approaches often do not work at scale [**V**]. Accordingly, a platform-agnostic, automated approach with good performance at scale was missing. In [**V**], we proposed an automated framework and evaluated the extent to which we can cluster digital ego networks based on their tie strength. We also analysed the structural and interactional characteristics that distinguish weak-tie networks from strong-tie ones. Finally, we analysed how tie-strength patterns vary across different regions and demographic groups in online networks.

We utilized four measures studied in detail in [**IV**] for clustering. Precisely, a) interaction strength (IS): a weighted interaction frequency across edges, b) relative interaction strength (RIS): proportionate of interaction happens between alters vs. ego-alter interactions, c) social similarity (SS): shared of common friends inside the network, and d) outliers (OUT) percentage of nodes become isolated if ego is removed from the network. We extracted these measures from each network and represented each network as a 4-dimensional feature vector.

After transforming our network dataset into a 4D vector space, we applied a cleaning and filtering phase to avoid noisy clustering. We defined thresholds for each measure and removed invalid cases. Also, to filter outlier datapoints, we applied an iterative two-threshold (2T) approach integrated with median absolute deviation (MAD) on the data [V, 84]. Next, we clustered our dataset into four clusters using repeated k-means with multiple initializations to avoid potential local optima. Finally, we labeled the final clusters by their centroid strength as weak, moderately-weak, moderately-strong, and strong.

To validate the number of clusters for tie strength, we used clustering metrics such as the Silhouette [85], Calinski–Harabasz [86], and WB-index [87]. We test these metrics across different numbers of clusters; the first two metrics indicate that k=4 is the best. However, in the case of web index, k=4 was the second-best option, with a slight difference from the best option, k=5.

As one part of the experiments, we compared cluster distributions across locations. Our results revealed a strong regional pattern [**V**]. For instance, Nordic networks demonstrated the highest share of weak-tie networks and the lowest share of strong-tie networks. Australian networks were the opposite of Nordic networks and had the highest share of strong-tie networks and the lowest share of weak-tie networks [**V**]. UK networks were more balanced than those in other locations, and the US network leaned more toward weak-tie networks [**V**].

Finally, we compared the distribution of tie strength across genders. The digital social network corpora we used in [**V**] were assigned gender labels by Fränti et al. [57]. Our results indicated that male networks (networks with male ego nodes) are more common across all four tie-strength categories than female networks (networks with female ego nodes). In addition, there was one overall pattern in the data: female and uncategorized accounts appeared relatively more in the weaker tie clusters than in the moderately-strong and strong clusters [**V**].

6 Summary of contributions

[**I**]: The paper aims to test whether network size affects the distinction between weak and strong ties in digital social networks and how this impacts language use and change. Most sociolinguistic network studies used small-scale ethnographic data. Consequently, there is limited empirical evidence on how network size affects the weak-tie vs. strong-tie distinction. This paper introduces a computational method for analyzing large Twitter-based social networks and demonstrates how this method can be utilized to investigate language variation and innovation at scale. Our results indicate that, first, at approximately 120 nodes, the difference between weak and strong ties disappears. Second, weak and strong ties contribute similarly to language diffusion in large networks. These findings contradict earlier theories based on small networks, where only weak ties were seen as channels for innovation. This paper extends sociolinguistic research into large digital environments, which is valuable for analyzing language change and information diffusion in large digital networks.

[**II**]: This paper investigates how well user interaction history, activity pattern, and generated content can predict perceived similarity between Twitter users. Previous studies mainly relied on subjective perceptions without identifying which type of user-generated data best predicts perceived similarity in social networks. We first introduce a dataset combining Twitter data with user-reported similarity perceptions. Second, three computational methods based on interaction, activity, and hashtag are proposed to measure similarity. Our results demonstrate that interaction-based similarity outperforms the other methods with the highest accuracy. In addition, the accuracy for measuring similarity decreases as the network size increases. Finally, male ego networks display slightly higher similarity detection accuracy than female ones across all methods. The findings offer a practical method for identifying similar users in social networks, useful for researchers in social network analysis and recommendation systems.

[**III**]: This paper analyzes social connections among Twitter users in the Nordic countries and examines whether users cluster by country based on friends networks. Previous studies on national Twitterspheres often rely on hashtags or sub-sampling, but there is limited large-scale research using graph clustering to analyze country-level patterns in multi-country regions like the Nordics. We created a large Twitter network across five Nordic countries utilizing geo-tagged data. Then, we applied a recent graph clustering algorithm (M-algorithm), evaluated how closely user clusters aligned with national borders, and analyzed clusters content by country. Our results illustrate that clusters align strongly with home countries; over 90% of user links are within the same country. Also, five distinct clusters corresponding to five Nordic countries were identified, with no clear evidence of sub-clusters within countries. Finland had the highest internal connection rate (99%), while Sweden had the most external links. Finally, hashtag use differed significantly between countries, supporting the clustering results. This paper provides insights for researchers interested in social media data by revealing strong national clustering in user interactions and content, demonstrating the effectiveness of graph clustering methods, and high-

lighting digital divides in geographically close regions.

[**IV**]: This paper aims to identify reliable measures for quantifying the strength of ties in digital ego networks and evaluate how these measures perform across real-world and random Twitter-based networks. Previous studies considered tie strength a unidimensional concept depending on measures such as interaction frequency or emotional closeness, which oversimplifies the concept. Consequently, a multidimensional approach that considers multiple factors for measuring tie strength was lacking. This study proposes and evaluates eight measures to calculate tie strength utilizing large-scale Twitter data. Our results demonstrate that out of the eight studied measures, Interaction Strength (IS), Social Similarity (SS), and Outliers (OUT) are the most effective at distinguishing the real-world from random networks and measuring tie strength. In addition, Nordic ego networks tend to have the weakest tie structures, especially based on IS and OUT measures. Researchers studying social networks, information flow, and language change can use the proposed approach to better model and simulate social behavior, such as how innovations, slang, or ideas spread depending on whether ties are weak or strong.

[**V**]: This paper explores how social media networks can be clustered into weak-tie and strong-tie networks based on measurable features. In addition, what are the statistical characteristics that differentiate these clusters? Measuring tie strength based on heuristics or thresholds that exist in the literature are not scalable or automated. Also, there is no comprehensive, data-driven method for clustering large-scale online networks by tie strength. Using k-means clustering and four tie strength measures, we developed a pipeline to group ego networks into weak, moderate, and strong-tie categories. We applied the method to a large geo-located Twitter dataset (DSN corpora), offering a scalable way to study tie strength in social media. Our results demonstrate that UK, US, and Australian users mostly form strong-tie networks. In addition, Nordic users are more likely to form weak-tie or moderate-tie networks. The findings can be utilized by marketers and advertisers targeting users based on the network type and platform designers improving user experiences and content delivery by understanding tie strengths.

7 Conclusion

Online social networks shape how people share information, innovate, use language, and build trust. They influence many aspects of daily life, yet they remain difficult to study systematically. Classic social theories were developed using observations of small, local networks, while contemporary born-digital data are massive, metadata-rich, and structurally complex. This thesis positions itself as a bridge between these two worlds: it connects foundational theories, particularly the weak-tie hypothesis and broader tie strength theory, with scalable computational models capable of analyzing large-scale social media networks.

Throughout this thesis, the ego network served as the central unit of analysis. The ego, its alters, the ego–alter and alter–alter connections, and the interactions occurring along these ties constituted the core structural and behavioral elements examined. Across the studies, our recurring objective is to move beyond manual, edge- or node-level analyses and single metric approaches. Instead, we advance toward platform-agnostic, network-level, multi-dimensional computational modelling capable of capturing the full complexity of social structures at scale.

We revisited the classic sociolinguistic claim that weak-tie networks facilitate linguistic innovation, whereas strong-tie networks reinforce norms and resist change. Earlier support for this theory is drawn primarily from small-scale ethnographic studies, raising questions about its applicability to large-scale social media networks. In this thesis, we examined whether computational methods can operationalize tie strength at scale using Twitter ego networks and whether the theoretical distinction between weak and strong environments holds as network size increases. To do so, we defined linguistic innovation using text-based markers and compared diffusion patterns across weak and strong tie environments. The results showed that network size plays a critical conditioning role: in smaller ego networks, weak and strong ties exhibit distinct diffusion behaviors, consistent with the classic theory. However, as networks grow, these differences diminish, and around a threshold of roughly 120 nodes, the distinction effectively disappears.

We analyzed user similarity in online environments to determine which digital signals best predict perceived similarity on social media. Although prior research acknowledges that perceived similarity between online users is inherently subjective, it remains unclear which observable online cues align most closely with human judgment. To address this, we collected ground-truth similarity assessments through an online survey and compared them with three computational approaches based on interaction patterns, activity profiles, and user-generated content. Our findings demonstrated that interaction-based similarity aligns most closely with human perceptions and consistently outperforms the other methods. In addition, we observed that the accuracy of similarity estimation is conditioned by network size: as networks grow larger, the ability to measure similarity accurately decreases, regardless of which computational approach is used.

Next, we examined community detection within the Nordic Twittersphere. While previous Twittersphere studies typically focused on a single region or language, a multicountry, multilingual analysis of this scale has been largely missing. To ad-

dress this gap, we constructed a geographically labeled network dataset representing the Nordic Twittersphere and applied the Malgorithm for clustering, comparing results across several cost functions. Our findings indicated that community structure aligns strongly with national borders: each Nordic country forms a distinct cluster, and no additional sixth regional-level cluster emerges. We also found no meaningful subclusters within individual country clusters. Furthermore, the similarity relationships between countries' content-based clusters do not mirror their connectivity patterns, indicating that content similarity and structural connections follow different logics in this context.

Moreover, unlike traditional research on computer-mediated communication, we analyzed social media data while explicitly accounting for its networked structure. To support this, we introduced the Network Strength Index (NSI), a network-level measure of tie strength constructed from eight indicators. The NSI maps each ego network onto a continuous scale from 0 to 1, enabling systematic ranking and comparison of networks by their overall tiestrength configuration. We also developed the Digital Social Network (DSN) corpora, a large geolabeled dataset comprising Twitter ego networks from Australia, the United Kingdom, the United States, and the Nordic region. Using this dataset, we validated the eight tiestrength indicators against randomly generated networks and evaluated their robustness across variations in size and degree. Our results showed that interaction strength, social similarity, and the outliers are the most informative and reliable measures for distinguishing real-world networks from random baselines. Furthermore, the DSN analysis revealed clear regional differences: weaktie networks are more prevalent in the Nordic region compared to the other three regions in the dataset.

Finally, we developed a platform-agnostic clustering pipeline to cluster online social networks based on their tiestrength profiles. Using the four best-performing indicators identified through the NSI framework, interaction strength, relative interaction strength, social similarity, and the outliers, we represented each ego network as a four-dimensional feature vector. We then applied repeated k-means clustering with random initialization to partition the networks into four ordered tiestrength categories: weak, moderately-weak, moderately-strong, and strong. After validating the number of clusters, we conducted analyses based on cluster composition and user gender. Our results showed that tiestrength patterns vary systematically across regions. Nordic networks are skewed toward weaker tiestrength categories, whereas Australian networks are skewed toward stronger ones. UK networks displayed a more balanced distribution, while US networks leaned toward the weaktie clusters. The gender analysis revealed that, across all tie-strength categories, male networks are the most common. In contrast, female and uncategorized networks are disproportionately represented in the weaker clusters and are less common in the strongest tiestrength category.

Across this thesis, three overarching themes connect the papers. First, network size consistently shapes what can be observed in the analyses and results. In [**I**], the behavioral distinction between weaktie and strongtie environments diminishes once ego networks exceed roughly 120 nodes. In [**II**], the accuracy of usersimilarity predictions declines as network size increases. In [**IV**], tiestrength indicators are explicitly tested for robustness with respect to size and degree, demonstrating how strongly network scale conditions measurement outcomes. Second, the thesis advances a shift from classic social-theoretical concepts, such as weak ties and community structure, toward scalable, operational tools for analyzing large online networks. This progression culminates in the development of the reusable network

strength index [**IV**] and a platform-agnostic clustering framework [**V**], both of which treat social media not merely as text streams but as complex networked environments. Third, the findings show that geography continues to play a meaningful role in online social structures. In [**III**], connectivity-based clusters in the Nordic Twittersphere align closely with national borders, and in [**V**], tiestrength distributions vary systematically across regions. Together, these themes demonstrate how structural scale, theoretical grounding, and spatial context jointly shape social processes in online networks.

Future work can extend this thesis in several clear directions. First, the NSI and the clustering framework should be applied beyond Twitter to assess how well they transfer to other directed online networks and whether their performance generalizes across platforms with different interaction norms. Second, improving location inference methods would substantially strengthen all regional and demographic analyses, as geographic signals remain incomplete and uneven across users and platforms. Third, the NSI model itself can be refined by further validating the constituent measures and adjusting the feature set or weighting scheme, given that some indicators are consistently more informative than others. Finally, the framework should be linked to a broader range of outcomes beyond the initial linguistic diffusion case study, and the size effect warrants direct investigation to better understand why the weak/strong tie distinction dissipates around 120 nodes and under what conditions that threshold may shift.

BIBLIOGRAPHY

[1] A. S. Z. Chase, A. Kamp-Whittaker, and M. A. Peeples, "Archaeologies of people and space: Social network analysis of communities and neighborhoods in spatial context," *Journal of Anthropological Archaeology* **75**, 101607 (2024), https://doi.org/10.1016/j.jaa.2024.101607.

[2] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, 1994), https://doi.org/10.1017/CBO9780511815478.

[3] K. Jordan, "From social networks to publishing platforms: A review of the history and scholarship of academic social network sites," *Frontiers in Digital Humanities* **6** (2019), https://doi.org/10.3389/fdigh.2019.00005.

[4] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," *Science* **323**, 892–895 (2009), https://doi.org/10.1126/science.1165821.

[5] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication* **13**, 210–230 (2007), https://doi.org/10.1111/j.1083-6101.2007.00393.x.

[6] M. Vacchiano, B. Hollstein, R. A. Settersten, and D. Spini, "Networked lives: Probing the influence of social networks on the life course," *Advances in Life Course Research* **59**, 100590 (2024), https://doi.org/10.1016/j.alcr.2024.100590.

[7] C.-W. Chang and S.-H. Chang, "The impact of digital disruption: Influences of digital media and social networks on forming digital natives' attitude," *Sage Open* **13**, 21582440231191741 (2023), https://doi.org/10.1177/21582440231191741.

[8] F. Li, M. Shi, and R. Feng, "Social media use and job choices: The mediating roles of work values and self-efficacy," *Frontiers in Psychology* **16** (2025), https://doi.org/10.3389/fpsyg.2025.1485663.

[9] P. Puri, G. Hassler, S. Katragadda, and A. Shenk, "Digital cloning of online social networks for language-sensitive agent-based modeling of misinformation spread," *PLOS ONE* **19**, e0304889 (2024), https://doi.org/10.1371/journal.pone.0304889.

[10] J. L. Pouwels, T. Araujo, W. van Atteveldt, M. Bachl, and P. M. Valkenburg, "Integrating communication science and computational methods to study content-based social media effects," *Communication Methods and Measures* **18**, 115–123 (2024), https://doi.org/10.1080/19312458.2023.2285766.

[11] D. Surjatmodjo, A. A. Unde, H. Cangara, and A. F. Sonni, "Information pandemic: A critical review of disinformation spread on social media and

its implications for state resilience," *Social Sciences* **13**, 418 (2024), https://doi.org/10.3390/socsci13080418.

[12] U. Sharma and J. Singh, "A comprehensive overview of fake news detection on social networks," *Social Network Analysis and Mining* **14**, 120 (2024), https://doi.org/10.1007/s13278-024-01280-3.

[13] M. Laitinen, P. Rautionaho, M. Fatemi, and M. Halonen, "Do we swear more with friends or with acquaintances? F#ck in social networks," *Lingua* **320**, 103931 (2025), https://doi.org/10.1016/j.lingua.2025.103931.

[14] R. Ahnert, S. E. Ahnert, C. N. Coleman, and S. B. Weingart, *The Network Turn: Changing Perspectives in the Humanities* (Cambridge University Press, 2020), https://doi.org/10.1017/9781108866804.

[15] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology* **78**, 1360–1380 (1973).

[16] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons* **53**, 59–68 (2010), https://doi.org/10.1016/j.bushor.2009.09.003.

[17] D. Ruths and J. Pfeffer, "Social media for large studies of behavior," *Science* **346**, 1063–1064 (2014), https://doi.org/10.1126/science.346.6213.1063.

[18] K. Rajkumar, G. Saint-Jacques, I. Bojinov, E. Brynjolfsson, and S. Aral, "A causal test of the strength of weak ties," *Science* **377**, 1304–1310 (2022), https://doi.org/10.1126/science.abl4476.

[19] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and Twitter social networks," in *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 4 (2010), p. Article 1, https://doi.org/10.1609/icwsm.v4i1.14021.

[20] 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*, Vol. 10046, E. Spiro and Y.-Y. Ahn, eds. (Springer International Publishing, 2016), pp. 41–57, https://doi.org/10.1007/978-3-319-47880-7_3.

[21] M. Del Tredici and R. Fernández, "The road to success: Assessing the fate of linguistic innovations in online communities," in *Proceedings of the 27th International Conference on Computational Linguistics* (2018), pp. 1591–1603, https://aclanthology.org/C18-1135/.

[22] X. Zheng, J. Han, and A. Sun, "A survey of location prediction on Twitter," *IEEE Transactions on Knowledge and Data Engineering* **30**, 1652–1671 (2018), https://doi.org/10.1109/TKDE.2018.2807840.

[23] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st International Conference on World Wide Web* (2012), pp. 519–528, https://doi.org/10.1145/2187836.2187907.

[24] P. S. Park, J. E. Blumenstock, and M. W. Macy, "The strength of long-range ties in population-scale social networks," *Science* **362**, 1410–1413 (2018), https://doi.org/10.1126/science.aau9735.

[25] M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology* **83**, 1420–1443 (1978).

[26] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of the 13th International Conference on World Wide Web* (2004), pp. 491–501, https://doi.org/10.1145/988672.988739.

[27] J. Zhu and D. Jurgens, "The structure of online social networks modulates the rate of lexical change," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2021), pp. 2201–2218, https://doi.org/10.18653/v1/2021.naacl-main.178.

[28] M. Laitinen and M. Fatemi, "Testing the weak-tie hypothesis with social media," in *Proceedings of the 11th Conference on Computer-Mediated Communication and Social Media Corpora* (2024), p. 46, https://shs.hal.science/halshs-04673776/file/241007_CMC_Proceedings_DOI.pdf.

[29] 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," *Proceedings of the National Academy of Sciences* **104**, 7332–7336 (2007), https://doi.org/10.1073/pnas.0610245104.

[30] K. Kucher, M. Fatemi, and M. Laitinen, "Towards visual sociolinguistic network analysis," (2021), pp. 248–255, https://doi=10.5220/0010328202480255.

[31] Q. Yao, R. Y. M. Li, L. Song, and M. J. C. Crabbe, "Construction safety knowledge sharing on Twitter: A social network analysis," *Safety Science* **143**, 105411 (2021), https://doi.org/10.1016/j.ssci.2021.105411.

[32] E. Kopacheva, M. Fatemi, and K. Kucher, "Using social-media-network ties for predicting intended protest participation in Russia," *Online Social Networks and Media* **37–38**, 100273 (2023), https://doi.org/10.1016/j.osnem.2023.100273.

[33] C. McMillan, "Worth the weight: Conceptualizing and measuring strong versus weak tie homophily," *Social Networks* **68**, 139–147 (2022), https://doi.org/10.1016/j.socnet.2021.06.003.

[34] M. Laitinen and M. Fatemi, "Big and rich social networks in computational sociolinguistics," in *Social and Regional Variation in World Englishes* (Routledge, 2022).

[35] M. E. Brashears and E. Quintane, "The weakness of tie strength," *Social Networks* **55**, 104–115 (2018), https://doi.org/10.1016/j.socnet.2018.05.010.

[36] D. Nguyen, A. S. Doğruöz, C. P. Rosé, and F. de Jong, "Computational sociolinguistics: A survey," *Computational Linguistics* **42**, 537–593 (2016), https://doi.org/10.1162/COLI_a_00258.

[37] J. Lundberg, J. Nordqvist, and M. Laitinen, "Towards a language independent Twitter bot detector," in *Proceedings of the Digital Humanities in the Nordic Region (DHN2019): 4th Conference of The Association Digital Humanities in the Nordic Countries* (2019).

[38] T. A. B. Snijders, G. G. van de Bunt, and C. E. G. Steglich, "Introduction to stochastic actor-based models for network dynamics," *Social Networks* **32**, 44–60 (2010), Dynamics of Social Networks. https://doi.org/10.1016/j.socnet.2009.02.004.

[39] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology* **27**, 415–444 (2001), https://doi.org/10.1146/annurev.soc.27.1.415.

[40] K. Z. Khanam, G. Srivastava, and V. Mago, "The homophily principle in social network analysis: A survey," *Multimedia Tools and Applications* **82**, 8811–8854 (2023), https://doi.org/10.1007/s11042-021-11857-1.

[41] M. Van Zalk and J. Denissen, "Idiosyncratic versus social consensus approaches to personality: Self-view, perceived, and peer-view similarity," *Journal of Personality and Social Psychology* **109**, 121–141 (2015), https://doi.org/10.1037/pspp0000035.

[42] M. Arazzi, M. Ferretti, S. Nicolazzo, and A. Nocera, "The role of social media on the evolution of companies: A Twitter analysis of Streaming Service Providers," *Online Social Networks and Media* **36**, 100251 (2023), https://doi.org/10.1016/j.osnem.2023.100251.

[43] M. Laitinen and M. Fatemi, "Data-intensive sociolinguistics using social media," *Annales Academiae Scientiarum Fennicae* **2023**, 38–61 (2023), https://doi.org/10.57048/aasf.136177.

[44] F. V. Münch, B. Thies, C. Puschmann, and A. Bruns, "Walking through Twitter: Sampling a language-based follow network of influential Twitter accounts," *Social Media + Society* **7**, 2056305120984475 (2021), https://doi.org/10.1177/2056305120984475.

[45] A. Bruns, B. Moon, F. Münch, and T. Sadkowsky, "The Australian Twittersphere in 2016: Mapping the follower/followee network," *Social Media + Society* **3**, 2056305117748162 (2017), https://doi.org/10.1177/2056305117748162.

[46] A. Bruns and G. Enli, "The Norwegian Twittersphere: Structure and dynamics," *Nordicom Review* **39**, 129–148 (2018), https://doi.org/10.2478/nor-2018-0006.

[47] S. Sieranoja and P. Fränti, "Adapting k-means for graph clustering," *Knowledge and Information Systems* **64**, 115–142 (2022), https://doi.org/10.1007/s10115-021-01623-y.

[48] J. Eisenstein, "Identifying regional dialects in on-line social media," in *The Handbook of Dialectology* (John Wiley & Sons, Ltd., 2017), pp. 368–383, https://doi.org/10.1002/9781118827628.ch21.

[49] D. Centola and M. Macy, "Complex contagions and the weakness of long ties," *American Journal of Sociology* **113**, 702–734 (2007), https://doi.org/10.1086/521848.

[50] K. B. Wright and C. H. Miller, "A measure of weak-tie/strong-tie support network preference," *Communication Monographs* **77**, 500–517 (2010), https://doi.org/10.1080/03637751.2010.502538.

[51] A. Bruns, "After the 'APIcalypse': Social media platforms and their fight against critical scholarly research," *Information, Communication & Society* **22**, 1544–1566 (2019), https://doi.org/10.1080/1369118X.2019.1637447.

[52] C. Montag, B. J. Hall, and Y.-H. Lin, "Is it possible to circumnavigate the APIcalypse? On challenges to study mental health in the age of digitalization and AI," *The European Journal of Psychiatry* **38**, 100273 (2024), https://doi.org/10.1016/j.ejpsy.2024.100273.

[53] R. I. M. Dunbar, *Grooming, Gossip, and the Evolution of Language* (Harvard University Press, 1996).

[54] V. Arnaboldi, A. Passarella, M. Conti, and R. I. M. Dunbar, "Chapter 5—Evolutionary dynamics in Twitter ego networks," in *Online Social Networks*, V. Arnaboldi, A. Passarella, M. Conti, and R. I. M. Dunbar, eds. (Elsevier, 2015), pp. 75–92, https://doi.org/10.1016/B978-0-12-803023-3.00005-9.

[55] A. Tabarcea, N. Gali, and P. Fränti, "Framework for location-aware search engine," *Journal of Location Based Services* **11**, 50–74 (2017), https://doi.org/10.1080/17489725.2017.1407001.

[56] A. Tabarcea, V. Hautamäki, and P. Fränti, "Ad-hoc georeferencing of webpages using street-name prefix trees," in *Web Information Systems and Technologies*, J. Filipe and J. Cordeiro, eds. (Springer, 2011), pp. 259–271, https://doi.org/10.1007/978-3-642-22810-0_19.

[57] 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 *12th International Conference on CMC and Social Media Corpora for the Humanities* (CMC-Corpora, 2025).

[58] N. Fazal, K. Q. Nqyuen, and P. Fränti, "Efficiency of web crawling for geotagged image retrieval," *Webology* **16**, 16–39 (2019).

[59] N. Fazal and P. Fränti, "Social media data for content creation in location-based games," *Journal of Location Based Services* **19**, 167–194 (2024), https://doi.org/10.1080/17489725.2024.2414000.

[60] L. Milroy, *Language and Social Networks* (Wiley, 1987), https://www.wiley.com/en-it/Language+and+Social+Networks%2C+2nd+Edition-p-9780631153146.

[61] L. Milroy and J. Milroy, "Social network and social class: Toward an integrated sociolinguistic model," *Language in Society* **21**, 1–26 (1992), https://doi.org/10.1017/S0047404500015013.

[62] E. W. Schneider, "Sociolinguistic theory: Linguistic variation and its social significance," *Journal of English Linguistics* **27**, 49–56 (1999), https://doi.org/10.1177/00754249922004426.

[63] C. McCarty, "Structure in personal networks," *The Journal of Social Structure* (2002), https://www.semanticscholar.org/paper/Structure-in-Personal-Networks-McCarty/856eba52da9200f18dabaed6ccd0399df108215d.

[64] P. V. Marsden, "Network data and measurement," *Annual Review of Sociology* **16**, 435–463 (1990), https://doi.org/10.1146/annurev.so.16.080190.002251.

[65] C. McCarty, P. D. Killworth, and J. Rennell, "Impact of methods for reducing respondent burden on personal network structural measures," *Social Networks* **29**, 300–315 (2007), Special Section: Advances in Exponential Random Graph (P*) Models. https://doi.org/10.1016/j.socnet.2006.12.005.

[66] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," in *Proceedings of the 19th International Conference on World Wide Web* (2010), pp. 591–600, WWW '10. https://doi.org/10.1145/1772690.1772751.

[67] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Computational social science," *Science* **323**, 721–723 (2009), https://doi.org/10.1126/science.1167742.

[68] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," *Social Networks* **25**, 211–230 (2003), https://doi.org/10.1016/S0378-8733(03)00009-1.

[69] N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the National Academy of Sciences* **106**, 15274–15278 (2009), A. Pentland (Sandy). https://doi.org/10.1073/pnas.0900282106.

[70] J. McLevey, J. Scott, and P. J. Carrington, eds., *The Sage Handbook of Social Network Analysis* (Sage Publications Ltd, 2024), https://doi.org/10.4135/9781529682618.

[71] P. Fränti and S. Sieranoja, "Clustering accuracy," *Applied Computing and Intelligence* **4**, 24–44 (2024), https://doi.org/10.3934/aci.2024003.

[72] M. Rezaei and P. Fränti, "K-sets and k-swaps algorithms for clustering sets," *Pattern Recognition* **139**, 109454 (2023), https://doi.org/10.1016/j.patcog.2023.109454.

[73] P. Fränti, S. Sieranoja, K. Wikström, and T. Laatikainen, "Clustering diagnoses from 58 million patient visits in Finland between 2015 and 2018," *JMIR Medical Informatics* **10**, e35422 (2022), https://doi.org/10.2196/35422.

[74] N. Strayer, S. Zhang, L. Yao, T. Vessels, C. A. Bejan, R. S. Hsi, J. K. Shirey-Rice, J. M. Balko, D. B. Johnson, E. J. Phillips, A. Bick, T. L. Edwards, D. R. Velez Edwards, J. M. Pulley, Q. S. Wells, M. R. Savona, N. J. Cox, D. M. Roden, D. M. Ruderfer, and Y. Xu, "Interactive network-based clustering and investigation of multimorbidity association matrices with associationSubgraphs," *Bioinformatics* **39**, btac780 (2023), https://doi.org/10.1093/bioinformatics/btac780.

[75] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," *Applied Intelligence* **48**, 4743–4759 (2018), https://doi.org/10.1007/s10489-018-1238-7.

[76] J. G. Montero, A. Karjus, K. Smith, and R. A. Blythe, "Reliable detection and quantification of selective forces in language change," *Corpus Linguistics and Linguistic Theory* **21**, 31–73 (2025), https://doi.org/10.1515/cllt-2023-0064.

[77] W. T. Fitch, "Empirical approaches to the study of language evolution," *Psychonomic Bulletin & Review* **24**, 3–33 (2017), https://doi.org/10.3758/s13423-017-1236-5.

[78] A. Tommasel and D. Godoy, "Influence and performance of user similarity metrics in followee prediction," *Journal of Information Science* **48**, 600–622 (2022), https://doi.org/10.1177/0165551520975359.

[79] A. From, E. Diamond, N. Kafaee, M. Reynaga, R. S. Edelstein, and A. M. Gordon, "Does similarity matter? A scoping review of perceived and actual similarity in romantic couples," *Journal of Social and Personal Relationships* **42**, 2780–2802 (2025), https://doi.org/10.1177/02654075251349720.

[80] R. M. Montoya, R. S. Horton, and J. Kirchner, "Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity," *Journal of Social and Personal Relationships* **25**, 889–922 (2008), https://doi.org/10.1177/0265407508096700.

[81] S. Peng, Y. Zhou, L. Cao, S. Yu, J. Niu, and W. Jia, "Influence analysis in social networks: A survey," *Journal of Network and Computer Applications* **106**, 17–32 (2018), https://doi.org/10.1016/j.jnca.2018.01.005.

[82] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Computing Surveys* **52**, 5:1–5:38 (2019), https://doi.org/10.1145/3285029.

[83] M. Laitinen, J. Lundberg, M. Levin, and R. Martins, "The Nordic tweet stream: A dynamic real-time monitor corpus of big and rich language data," https://erepo.uef.fi/handle/123456789/6697 (2018).

[84] J. Yang, S. Rahardja, and P. Fränti, "Outlier detection: How to threshold outlier scores?," in *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing* (2019), pp. 1–6, AIIPCC '19. https://doi.org/10.1145/3371425.3371427.

[85] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987), https://doi.org/10.1016/0377-0427(87)90125-7.

[86] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics* **3**, 1–27 (1974), https://doi.org/10.1080/03610927408827101.

[87] Q. Zhao and P. Fränti, "WB-index: A sum-of-squares based index for cluster validity," *Data & Knowledge Engineering* **92**, 77–89 (2014), https://doi.org/10.1016/j.datak.2014.07.008.

Paper **I**

Laitinen, M., **Fatemi, M.**, & Lundberg, J. (2020)
"Size matters: Digital social networks and language change"
Frontiers in Artificial Intelligence, 3
https://doi.org/10.3389/frai.2020.00046

Copyright statement

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 02 July 2020
doi: 10.3389/frai.2020.00046

Size Matters: Digital Social Networks and Language Change

Mikko Laitinen[1,2]*, *Masoud Fatemi*[1,2] *and Jonas Lundberg*[2]

[1] School of Humanities/English, University of Eastern Finland, Kuopio/Joensuu, Finland, [2] Center for Data Intensive Sciences and Applications, Linnaeus University, Växjö, Sweden

Social networks play a role in language variation and change, and the social network theory has offered a powerful tool in modeling innovation diffusion. Networks are characterized by ties of varying strength which influence how novel information is accessed. It is widely held that weak-ties promote change, whereas strong ties lead to norm-enforcing communities that resist change. However, the model is primarily suited to investigate small ego networks, and its predictive power remains to be tested in large digital networks of mobile individuals. This article revisits the social network model in sociolinguistics and investigates network size as a crucial component in the theory. We specifically concentrate on whether the distinction between weak and strong ties levels in large networks over 100 nodes. The article presents two computational methods that can handle large and messy social media data and render them usable for analyzing networks, thus expanding the empirical and methodological basis from small-scale ethnographic observations. The first method aims to uncover broad quantitative patterns in data and utilizes a cohort-based approach to network size. The second is an algorithm-based approach that uses mutual interaction parameters on Twitter. Our results gained from both methods suggest that network size plays a role, and that the distinction between weak ties and slightly stronger ties levels out once the network size grows beyond roughly 120 nodes. This finding is closely similar to the findings in other fields of the study of social networks and calls for new research avenues in computational sociolinguistics.

Keywords: social networks, Twitter, bot exclusion, data mining, weak ties, social network size

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Annibale Elia,
University of Salerno, Italy
Xavier Blanco Escoda,
Autonomous University of Barcelona, Spain

***Correspondence:**
Mikko Laitinen
mikko.laitinen@uef.fi

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *09 March 2020*
Accepted: *27 May 2020*
Published: *02 July 2020*

Citation:
Laitinen M, Fatemi M and Lundberg J (2020) Size Matters: Digital Social Networks and Language Change. Front. Artif. Intell. 3:46. doi: 10.3389/frai.2020.00046

INTRODUCTION

This article focuses on social networks and explores network size as a key determinant in the network theory used in sociolinguistics. Building on Granovetter (1973), the theory postulates that individuals form personal communities that provide a meaningful framework for them in their daily life (Milroy and Llamas, 2013). An individual's social network is the sum of relationships contracted with others, and a network may be characterized by ties of varying strength. If ties are strong and multiplex, the network is dense, and individuals are linked through close ties (such as friends). Conversely, ties can be weak in which case individuals are predominantly linked through occasional and insignificant ties (such as acquaintances), and the network is loosely knit. Most importantly, networks contribute to language maintenance and change. Ample empirical evidence shows that loose-knit networks promote innovation diffusion, whereas dense multiplex networks lead to communities that resist change (Milroy and Milroy, 1978, 1985; Milroy, 1987; Lippi-Green, 1989). The underlying reason for the weakness of strong ties in transmitting

innovation is the fear of losing one's social standing in a network. Adopting new ideas is socially risky, and we do not want to "rock the boat" in dense social structures.

Even though the social network theory is influential in sociolinguistics, it is mostly based on small data. Most studies have focused on what are usually referred to as ego networks obtained using ethnographic observation. According to Milroy and Milroy (1992, p. 5) this "effectively limits the field of study, generally to something between 30 and 50 individuals." Moreover, it has been suggested that the quantitative variable of a network "cannot be easily operationalized in situations where the population is socially and/or geographically mobile" (Milroy, 1992, p. 177). In this paper, we concentrate on networks that are larger than small networks of only a few dozen of individuals. This has been done because evidence from social anthropology suggests that average human networks are substantially larger, and individuals can maintain networks with well over 200 nodes (McCarty et al., 2001). Prior empirical work in sociolinguistics has therefore covered only a limited section of possible network sizes.

We have two research foci. First, we test the extent to which social media data from Twitter and computational methods could be utilized to operationalize network ties of highly mobile individuals in very large datasets. Second, we specifically concentrate on the effect of network size on the validity of the theory. We investigate if the fear of losing one's social standing by "rocking the boat" disappears in large strong-tie networks.

To respond to these questions, we discuss two computational methods that can take up large and messy social media data and render them usable for analyzing networks in sociolinguistics, thus expanding the empirical basis from small-scale ethnographic observations. The first method aims at uncovering broad quantitative patterns in data and utilizes what we call a cohort-based method of network size. The second consists of an algorithm-based approach that uses mutual interaction parameters in Twitter and aims to verify the patterns obtained using the cohort-based approach.

By doing so, the article continues our pilot investigation, which suggests that network size is a crucial component in the theory. Our first results indicated that weak ties are meaningful in small networks, but the distinction between truly weak ties and slightly stronger ties levels out when network size increases beyond a certain threshold level (Laitinen et al., 2017). This pilot was based on social media data that had not yet been cleaned of unwanted software robot data (i.e., bots). In the present study, we attempt to replicate the study using a more accurate dataset from which we have removed bots by means of machine-learning techniques and by using novel computational methods to test our first observations. Bot content can result in inaccuracies, and previous computational sociolinguistic studies rely on a range of methods when bots are handled. Their presence may be recognized, but they are nevertheless included in the results (Huang et al., 2016; Laitinen et al., 2017). Other methods, such as excluding material by using metadata parameters, are occasionally used (Coats, 2017), but as we demonstrate below in section Material and Methods, more advanced solutions are available.

As shown in the next section, the role of network size in sociolinguistics is an understudied phenomenon, which not only requires new tools but could also shed light on the contrast between strong and weak ties in innovation diffusion. One example is that while the weak-tie model is beneficial, it has recently seen substantial theoretical elaboration, and recent advances have broadened the understanding of networks ties as a unidimensional concept (Aral and Van Alstyne, 2011). What is clear is that weak-tie and close-knit networks are different for small ego networks obtained through ethnographic methods, but if network size is ignored, the social network theory is not fully consistent with some of the major findings in sociolinguistics. First, it is widely held that there is one period when individuals maintain maximally close ties with their peers, and that is adolescence (Chambers, 2003, p. 90–91). Yet, the role of adolescents in language change is indisputable and verified in both real-time and apparent-time studies of change in progress (Labov, 2001, p. 76; Tagliamonte and D'Arcy, 2009). There might, of course, be other reasons than interpersonal ties during adolescence that lead teens to diverge from adult norms, but network size deserves to be studied in more detail. Moreover, ample macro-level evidence suggests that densely populated and sufficiently large working-class urban areas have, throughout history, been sites for innovations (e.g., the Jewish quarters all over Europe, Harlem in New York City, or St. John's Ward in Toronto). Pan et al. (2013) suggest that it is the size and density of the ties of a center that are crucial for information diffusion. They investigate social-tie density and information contagion in urban populations, and their quantitative model shows how density, with both weak and strong ties, drives the "super-linear" growth of interaction and information diffusion. Close-knit urban centers may, of course, be sufficiently large to sustain individuals with weak ties through whom innovations spread to a community, but we simply do not yet know whether the role of weak and strong ties levels out beyond a certain threshold level.

Section Social Networks in Variationist Sociolinguistics Discusses not only the theoretical basis of social networks in sociolinguistics but also reviews recent insight from complex systems analysis and social network theory. Section Material and Methods details the material and the two methodologies. Section Results presents the results, and, finally, section Conclusions discusses the implications of our findings.

SOCIAL NETWORKS IN VARIATIONIST SOCIOLINGUISTICS

Social network analysis in the variationist paradigm transpires from the idea that individuals establish interpersonal ties of varying strengths to form communities. These personal social networks are not independent from other socio-cultural frameworks but are closely related to other variables, such as gender and social layer (Milroy and Milroy, 1992). Interpersonal ties influence the rate at which innovations are adopted and how they diffuse into a community. Sociolinguists have shown that strong networks tend to maintain and support local norms

and provide resistance to the adoption of competing norms from the outside. Conversely, conditions that are characterized by weak and uniplex ties are important channels for outside influence as people in such situations tend to accommodate to each other linguistically. Contact situations with weak ties therefore contribute positively to the spread of innovations.

This finding builds on Granovetter's (1973, p. 1365) observation that "only weak ties may be local bridges." More people can be reached through weak ties, but not all weak ties serve this function, "only those acting as bridges between network segments" (1983, p. 229). To explain this somewhat counterintuitive observation, Granovetter (1973, 1983) argues that close-knit networks encourage local cohesion and norm-enforcing communities where the adoption of innovations is risky. Loose-knit networks with individuals already on the social fringes are more susceptible to external innovations. In addition, weak ties may be expected to be more numerous among mobile individuals and are thus more likely to contribute to the diffusion of an innovation.

In variationist sociolinguistics, network ties have been operationalized in various ways (Milroy and Llamas, 2013). In the Belfast study, they were measured using five indicators to establish how complex and dense a particular tie was. The indicators consisted of (a) having membership in a locally-based group, (b) having ties with at least two households in the neighborhood, (c) sharing a workplace with two or more individuals from the neighborhood, (d) sharing a workplace with same-sex individuals from the neighborhood, and (e) being involved in voluntary activities with individuals from the same workplace. The responses resulted in a network strength scale, which formed an independent variable, and these values were compared to the dependent (phonological) variables. The results show that the individuals with strong network ties with the local community also exhibited the highest share of local, vernacular speech, and "that a close-knit network has an intrinsic capacity to function as a norm-enforcement mechanism, to the extent that it operates in opposition to larger scale institutional standardizing pressures" (Milroy and Milroy, 1985, p. 359).

A large body of variationist sociolinguistic literature exists in which the network-based approach has been applied to small contemporary communities (Milroy and Llamas, 2013). Milroy and Milroy (1978) use 46 speakers from three urban, blue-collar Belfast communities, and the network ties were established through a participant observation process in which a researcher was introduced to a community by means of a friend-of-a-friend technique. Of these, 12 had network scores qualifying them as weak tie individuals. The same also applies to Granovetter's (1973, p. 1368–1371) study as his empirical data came from a random sample of 100 personal interviews taken from the total sample of 282. Carefully constructed personal networks are obviously important, but the availability of social media data also forces us to ask if the model holds when tested with considerably larger networks.

Network size has not been considered as a separate independent variable in variationist sociolinguistics (Milroy and Llamas, 2013). The model has been applied to large communities in macro-level approaches (Milroy and Milroy, 1985; contrasting Icelandic and English; Raumolin-Brunberg, 1996; investigating mobility as a result of the Civil War in the seventeenth-century England, and Nevalainen, 2000; examining patterns of mobility in Early Modern London). However, while all of these studies are rich in linguistic evidence, they nevertheless contain no direct quantitative evidence of how much weak ties actually increase in the settings that are examined. They rely on indirect evidence of migration patterns, population growth and birth/death rates for instance, but information of average network size per community is not detailed.

Recent findings in social anthropology have shown that an average network size is larger than a few dozen individuals. Dunbar (1992, p. 469) has suggested that the neocortex size and the number of neocortical neurons impose a cognitive upper limit on an individual's information-processing capacity. These limit "the number of relationships that an individual can monitor simultaneously" to around 150 nodes. Additionally, McCarty et al. (2001) use two methods to estimate the size of average networks. They use what they term the scale-up and summation methods, and the results show "a remarkable similarity between the average network size[s] generated by both methods (∼291)" (2001, p. 28). They estimate, however, that network sizes for various subpopulations can be substantially larger. These include clergy, politicians, labor organizers, and diplomats.

Sociolinguistic research has covered a part of the feasible network sizes. **Figure 1** visualizes this with the aid of dummy data. The x-axis indicates the size of networks and the y-axis the rate of innovation adoption for network types. The left-hand part shows the size of the networks covered, while the right shows how these fare with cognitively possible human network sizes.

We added a regression line to the visualizations but given the absence of empirical evidence it is impossible to know whether the line continues if we have evidence exclusively from small networks.

Recent findings from fields outside sociolinguistics suggest that network sizes play a more substantial role than previously thought. Ma et al. (2019) focus on trust in public and private social media groups, surveying 6,383 Facebook Groups users. Their observations show that people trust private groups more than they do public groups, which is to be expected. However, the differences between group types disappear once the group size exceeds circa 150 members. When networks become larger, individuals are no longer be able to perform the mental reasoning of who actually is in the group and who is not. Therefore, the difference between network types levels in large networks.

Moreover, increasing empirical evidence has recently led social network scholars to question the unidimensionality of the weak-tie model. Brashears and Quintane (2018) for instance elaborate on the idea of bandwidth in social contacts as an additional dimension. This concept refers to the total flow of information and accounts for capacity, frequency, and redundancy of network ties. Their model shows that even though humans acquire a smaller proportion of new ideas through strong contacts, the greater bandwidth of these contacts means that more total content is transmitted through these contacts. Strong contacts could therefore be more likely to transmit a greater share of novel information than weak ties, which could

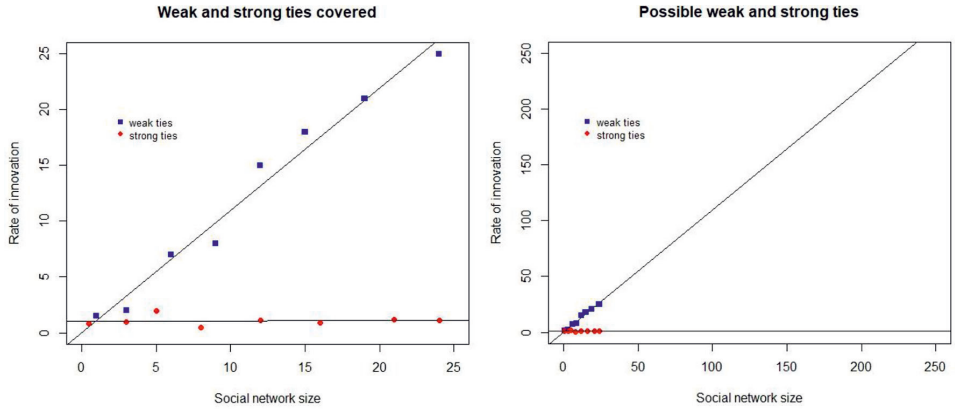


FIGURE 1 | A schematic representation of the network sizes covered (**Left**) and also the cognitively possible networks (**Right**).

explain the role of large urban working-class centers as places for innovation.

We investigate networks in Twitter and operationalize them using metadata available for each account. These are related to network size and mutual interaction patterns. Previous studies in computational sociolinguistics have used such information more to extract social network structures (Nguyen et al., 2016), but less to deepen understanding of the social network theory, which is the locus of this article. Eleta and Golbeck (2014) apply machine learning to study how social network characteristics and linguistic profiles influence language choice and how multilingual users of Twitter mediate between language groups in their social networks Their data consist of 92 ego networks, and the observations show that the proportion of English users in the network is the most significant predictor of language choice. Moreover, if a network consists of L2 users, this will increase the likelihood of L2 use. Kim et al. (2014) investigate how virtual networks impact multilingual practices, and they quantify "the degree to which users are the 'bridge-builders' between monolingual language groups." Hale (2014) studies networks utilizing mentions and retweets, and his results confirm the central role of multilingual users, and those who use English in particular, as the bridging forces in the network.

MATERIALS AND METHODS

To test the computational methods, we use two sets of Twitter data. Section A Cohort-Based Approach to Network Size uses evidence from the *Nordic Tweet Stream* corpus (NTS), which is a real-time monitor corpus of geolocated tweets and their metadata from the five Nordic nations (Laitinen et al., 2018). Section An Algorithmic Approach to Networks in Sociolinguistics utilizes an algorithm-based method, which makes use of mutual interaction data from a set of accounts from the Nordic region.

The NTS is being collected using the free Twitter Streaming API and the HBC (https://github.com/twitter/hbc) as the downloading mechanism. We apply a double filtering with the geolocation information and the Nordic country codes to ensure that the material originates from the region (Laitinen et al., 2018). While tweet data offer an efficient way of capturing big societal data, there are limitations. As an illustration, users who do not want to share their geolocation are not included. Depending on privacy settings and the geolocation method used, tweets either have (a) an exact location specified as a pair of latitude and longitude coordinates or (b) an approximate location specified as a rectangular bounding box. These geolocation data are available in the metadata attached to the message. Alternatively, no location at all is specified. For location, the data are derived either from the user's device itself (using the GPS) or by detecting the location of the user's Internet Protocol (IP) address (GeoIP). Exact coordinates are almost certainly from devices with built-in GPS receivers (e.g., phones and tablets). The GeoIP-based device location can be tricked by using proxy gateways. Attempting to hide one's location is probably most common amongst users with a malicious intent, such as bots.

To exclude bots and to increase data accuracy, we use a machine-learning algorithm developed by Lundberg et al. (2019). The version recognizes automatically generated tweets (AGTs) written in English and in Swedish. We define an AGT as a tweet in which all or parts of the natural language content are generated automatically by a bot or other type of program. The algorithm makes use of nine numerical and nominal properties that can be computed directly from the tweet metadata. The accuracy rate of the algorithm is over 97%. The results in section A Cohort-Based Approach to Network Size exclude possible bot accounts, whose share of AGTs is $>50\%$, and section An Algorithmic Approach to Networks in Sociolinguistics focuses on genuine human accounts that have been selected manually.

The first method (based on cohorts) does not assume a pre-existing social network as the starting point but rather aims at

TABLE 1 | Raw statistics for the data used in section An Algorithmic Approach to Networks in Sociolinguistics.

Account	Friends	Net size	Loss rate (%)	Tweets	Retrieval (in mins)	Text collection (in mins)
account_01	409	221	46	312,350	230	38
account_02	335	166	51	253,758	181	33
account_03	309	195	37	286,945	201	33
account_04	332	175	47	150,774	184	25
account_05	201	105	48	100,915	105	14
account_06	418	132	68	192,944	140	23
account_07	468	281	40	316,944	291	41
account_08	448	286	36	322,566	303	40
account_09	418	216	48	189,628	229	26
account_10	496	297	40	516,686	282	67

uncovering quantitative patterns in the data. To measure network sizes and to correlate size with the rate of innovation, we use two metadata attributes available for each tweet. They measure the number of one's online friends and followers, and networks are operationalized as follows: The number of followers indexes truly weak ties (i.e., requires no action from a user), and the number of friends is an indication of slightly stronger links (i.e., requires user effort). We suggested previously that these metadata offer a way of measuring social networks and are ideal for research purposes, because they are automatically generated and hence they reduce the observer bias (Laitinen et al., 2017). They are also freely available to researchers with intermediate computing skills.

Similar to Milroy (1987), we operate under the assumption that social networks are abstractions, but we also propose that information from digital social network applications can be used to distinguish between ties of varying strengths. Friend and follower counts are useful indicators of social networks because of their differing qualities. Our definition of truly weak ties and slightly stronger ties is similar to Granovetter's (1973, p. 1361) assumption that the "strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie." His methodology assumed stronger ties to be "friends," while weak ties consisted of "acquaintances," very similar to what we do below. By the same token, while we do not claim that friend count would indicate stronger ties in the sense in Milroy (1987), we assume that our operationalization of digital social networks is closely similar to the underlying idea of networks. Indeed, Milroy (1992, p. 178) argues that "a tie is 'weak' if it is less strong than the other ties against which it is measured," which also holds true for the follower counts when compared with friends.

The second method, the algorithm-based approach, zooms in on a set of real networks extracted by accessing account information through the Twitter API. We employ data such as friends and follower patterns, re-tweets, mentions, and directed messages. The accounts are anonymized, and we work with two types of network.

- Large (100–300 nodes) weak-tie networks
- Large (100–300 nodes) close-tie networks

We identified a set of accounts similar user profiles and extracted all interaction data available. The policy limitation of the API allows accessing up to 3,200 of the latest messages for each unprotected account. The account holders are from the metropolitan areas of Helsinki and Stockholm, are not working in academia, identify as males, have >10 messages primarily in English, and have more than 300 friends and followers. The last figure comes from a study that estimates median network sizes for multilingual individuals (Laitinen and Lundberg, 2020).

We narrowed the candidate accounts to ten and extracted their networks, including recent tweets and mutual interaction profiles. We excluded verified accounts (i.e., subpopulations with anomalous networks of politicians/celebrities/businesses) and accounts with more than 1,500 contacts (friends + followers). This was done to ease the time required for extracting mutual interaction data from large social networks. It is important to note that, while the number of accounts is small, the data extraction through the API takes circa 3–6 h per account (**Table 1**).

Even though the algorithm-based approach is tested with ten accounts, the size of our data is large. For instance, the mean network size is over 200 individuals (207), and the size of the textual data is over 2.6 million messages. In **Table 1**, the net size represents the number of collected accounts for the network (number of nodes in the graph). The loss ratio indicates the percentage of accounts lost after filtering.

The mutual interaction patterns are subjected to algorithms in order to assign labels of weak or strong networks to the accounts. The algorithms are explained in detail below, but they are mainly from the graph theory and the set theory, and some of them have been developed by us. For instance, we use betweenness centrality, which is a measure based on finding the shortest path between nodes (Freeman, 1977; Brandes, 2001) and closeness centrality (Perez and Germon, 2016). Kuikka (2018) argues that betweenness measures identify nodes that act as brokers between communities and are used to detect the density of how people are connected to each other in a network. We also use Jaccard Similarity Coefficient (JSC), which is a symmetric measure that calculates the similarity between two sets, and it is used to

measure the similarity between accounts in terms of the number of common followers/friends. The assumption is that the share of common friends/followers is higher in a strong-tie network than in weak-tie settings. In addition, we assign weights to each account in the network and employ a method which we call disjointness. This last method enables us to estimate how well the nodes in a network are connected if the ego node were to be removed. The network labels are therefore multidimensional.

As for the dependent variables, we employ items that are frequent enough to be used in the testing phase. First, the cohort-based method uses the dominant language for each account. This information is available in the NTS metadata, and the share of English messages per account is correlated with network sizes. As our data come from the Nordic region, it ought to be noted that while English has no *de jure* position in the region, it is increasingly used as a lingua franca. Space does not permit us to discuss the sociolinguistic diversity of the region, but see country overviews in Modiano (2003), Preisler (2003), Leppänen et al. (2011), and Graedler (2014). Previous studies that use Twitter data have suggested that a great majority of messages in one location, a region for instance, are from residents of that location (Gonçalves et al., 2018; Lamanna et al., 2018) and not from visitors and tourists. The cohort-based method uses information from tens of thousands of accounts, and we assume that our dataset is reliable, given the general limitations of Twitter data. We use automatically-assigned language labels, and although automated language identification methods are not entire accurate, the agreement between human coders and Twitter's language recognition system is fairly high for languages written in the Latin alphabet (Graham et al., 2013).

Second, the algorithmic approach uses a mixture of linguistic features available in the tweet text. These features consist of contracted forms (*won't*, *'ll*, *I'm* etc.), and *NEED* to used as a semi-modal auxiliary. These features are qualitatively different as the contracted forms index colloquial, spoken-like use (Biber et al., 1999, p. 1128–1132), while *NEED to* is currently undergoing change in English (Leech, 2013) and is highly pervasive in ELF use in the Nordic region (Laitinen, 2016).

RESULTS

A Cohort-Based Approach to Network Size

We illustrate the cohort-based method first using data from 199,832 accounts from the NTS, from which we removed subpopulations with anomalous network profiles, as defined in section Social Networks in Variationist Sociolinguistics. After the initial results, we test the findings with data from which software bots are removed. These bot-free data consist of 90,887 accounts, obtained from the NTS but limited to Sweden only (labeled as NTS-Human-Swe).

The null hypothesis is that increasing the number of network ties does not lead to increases in the share of English per account. The cohort-based approach for both categories is specified in (1)–(6) (it refers to followers in the NTS, but the same procedure applies to friends and to both datasets):

(1) We sort out all the accounts based on their followers' counts.

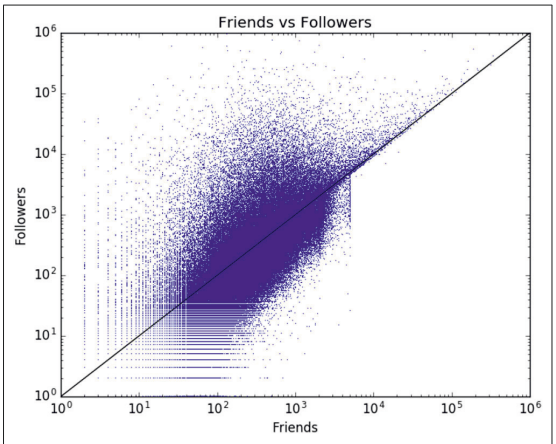


FIGURE 2 | Friends and followers visualized (199,832 accounts).

(2) The accounts are divided into N equally-sized cohorts where cohort 1 is the 199,832/N, and it has the lowest follower count, and cohort N has the highest. N can of course be adjusted.
(3) We compute the percentage of tweets written in English per each account.
(4) The language identifier used is Twitter's own language identification tool, the accuracy of which is discussed in section Material and Methods.
(5) We can adjust the proportions of English in the tweet stream (EngMajor) for each cohort and associate the cohorts with the EngMajor percentage. The results here use >50% share of messages in English (for other proportions, see Laitinen et al., 2017).
(6) We correlate the cohorts against the percentages and visualize them.

An average account profile in the NTS is such that the median size of networks is 235 friends and 195 followers. **Figure 2** shows how the friend and follower counts are distributed in the data. There is a relatively straightforward (x = y) spread of the values. The only exception is the friends category, in which Twitter imposes an upper limit of 5,000 friends that each individual account can follow (https://support.twitter.com/articles/66885#). The only way to increase one's friends count is to gain new followers, and therefore there is an even more direct correlation of friends/followers after the 5,000 mark.

Figure 3 (left) illustrates a 10-cohort division visualizing how cohorts differ in terms of the >50% percent threshold. The result shows that more Twitter followers means more messages in English, with the non-parametric Kendall tau correlation coefficient (0.956) indicating a strong positive correlation between the two vectors at statistically significant levels ($p < 0.0001$). Note that cohorts 1–4 are accounts with fewer than the median number of followers (i.e., 195).

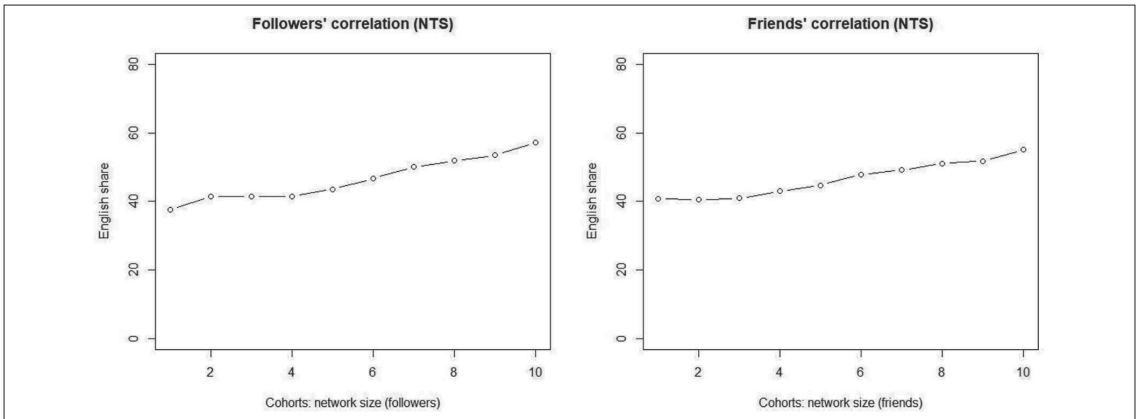


FIGURE 3 | The correlation between followers (**Left**) and friends (**Right**) and the share of accounts in which English dominates.

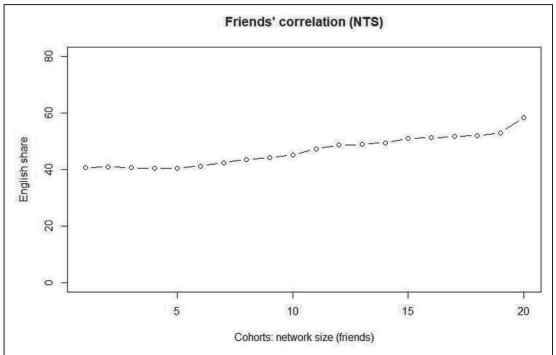


FIGURE 4 | The correlation between 20-cohort friend category and the EngMajor.

The quantitative pattern with these truly weak ties is clear. The correlation between the follower counts and the use of English is linear, and the correlation is strong. Of the accounts in which the number of followers is lower than the median, roughly 40% have the majority of their messages in English. The higher that we move in the cohorts, the higher is the share of such accounts. At the other extreme, in the cohorts with the highest number of followers over half of the accounts fulfill the criterion.

The quantitative pattern for the slightly stronger ties (friends) is shown on the right. The correlation between the number of friends and the increase in the use of English is strongly positive, with the Kendall tau correlation coefficient at 0.867 ($p < 0.0001$), i.e., for all of the 199,832 accounts in the dataset, more online friends means a larger share of messages in English.

However, contrary to what is observed with truly weak ties, the stronger network index behaves differently. For small networks, the increase in network size has no impact on the response variable. It is only from cohort 4 onwards that the share of EngMajor increases when we increase the number of friends in the network.

These results suggest that there is a straightforward correlation in the truly weak tie networks, but the friend data indicates that the distinction between weak ties and stronger ties levels out when the network size is large enough. If we had restricted our analysis only to traditional small networks of 30–50 nodes in ethnographic attempts, our data would have confirmed the customary finding related to the diffusion of innovations and network strength. That is, weak ties promote change, and stronger ties prevent it. However, the results obtained using this approach suggest that this is not necessarily the case. Once the network size grows to become large, the traditional distinction between weak and stronger ties disappears. Note that we are not referring to the percentages of the accounts, but to correlational patterns of the variable. Large networks here mean that the network sizes are still within the cognitive limits (see section Material and Methods).

To explain this finding, we must balance between the limitations and the advantages of our data. The most obvious limitation is that we might observe a random quantitative pattern that emerges from messy data. Moreover, we do not know anything about the density or the multiplexity of the network ties but can only assume that the friends category represents a slightly stronger network index, since it involves an active decision to follow someone. The friends network index might also include a greater share of interactive networks. To tackle the limitations, the next section applies a different method and approaches ego networks.

The obvious advantage is the size of our data. Each cohort in **Figure 3** consists of nearly 20,000 accounts, and we are not restricted to small ethnographic records. The network size for the first three cohorts is 0–122. As pointed out earlier, the median number of friends is 235. The results support rejecting the null hypothesis, but the threshold level of 122 stems from an arbitrary value of ten cohorts.

Figure 4 tests the observations using 20 cohorts. As the interest is on slightly stronger ties, we only use the friends data. The figure confirms the observation and indicates a leveled proportion of EngMajor for the first five cohorts. After that, the network size correlates positively with the increasing use of English in the tweet stream. The Kendall tau correlation coefficient is 0.905 at a statistically highly significant level ($p < 0.0001$).

Cohorts 1–5 consist of networks of <100 individuals, and a marked increase takes place only after cohort 5 (100–122 individuals). The share of accounts with a >50%+ share of tweets in English increases systematically for each cohort so that for cohort 6 it is 41.2%, and for cohort 19 it is 51.9%. Cohort 20 has its friends count at over 1700, and according to our present understanding, these represent "evangelists" in the Krishnamurthy et al. (2008) sense, i.e., they are more or less automated bots aiming at increasing their friends basis automatically.

Figure 4 suggests that the threshold network size after which the distinction between weak ties and slightly stronger ties levels is of around 122 nodes. Next, we zoom into the bot-free data, and the main question is whether we can replicate the findings using the bot-free data. Overall, the number of bots in the Swedish subset is low (1,149 accounts = 1.0%), but they generate a high number of tweets (404,804 = 7.6%). The majority language in the bots is English, since nearly 20% of all of the English tweets were identified as AGTs, but the corresponding share for Swedish was <2% (see Laitinen and Lundberg, 2020). The visualizations also exclude the smallest networks of fewer than five nodes.

The bot-free quantitative patterns are shown in **Figure 5**, and they are similar to those observed earlier. As for followers (left), they show a linear increase in the share of messages in the English per cohort as we move to the right on the x-axis. The correlation between network size and the share of English is not only straightforward but also statistically significant, as the Kendal tau correlation coefficient is one ($p < 0.0001$). For smaller networks, the share of English is around 40%, and it increases for every increase in the network size, so that the share for the largest networks is well over 50%. The increases are slight, but the shares of the English use nevertheless increase for each cohort. Once the network size grows larger, we observe more noteworthy increases.

The right-hand side visualizes the slightly stronger ties (friends) and verifies the initial observations. These results confirm the findings presented above. The observations show that the correlation with slightly stronger ties is equally linear, and this is also supported by the Kendal tau value (0.944, $p < 0.0001$). However, the share of English actually decreases for the small stronger-tie networks. That is, the empirical evidence presented here suggests that truly weak ties and slightly stronger ties behave slightly differently for small networks, but the distinction disappears once the network size grows larger. The share of English remains flat for cohorts 1–3 of the truly weak ties (left), while the share actually decreases for the slightly stronger ties for the smallest networks (right). Cohort 4 consists of those whose network size exceeds 120 nodes.

The present section has presented our cohort-based approach to measuring networks in social media. While we acknowledge that the method is straightforward, it has obvious benefits for this type of big and rich data approaches to language variability and social networks. The method is light in terms of computing power, as the values can be easily obtained from the data stream. In addition, we can use data in their entirety since each account makes the values directly available with minimal or no data loss.

The obvious difference between this approach and the ethnographically-oriented data-collection in Milroy (1987) is that our method does not deal with ego networks but rather takes a top-down approach, correlating network size and a linguistic feature. As for the innovation, previous studies have shown that English in the Nordic region is closely associated with age; this means that the younger generations clearly use English as an additional tool more often than do the older groups (Leppänen

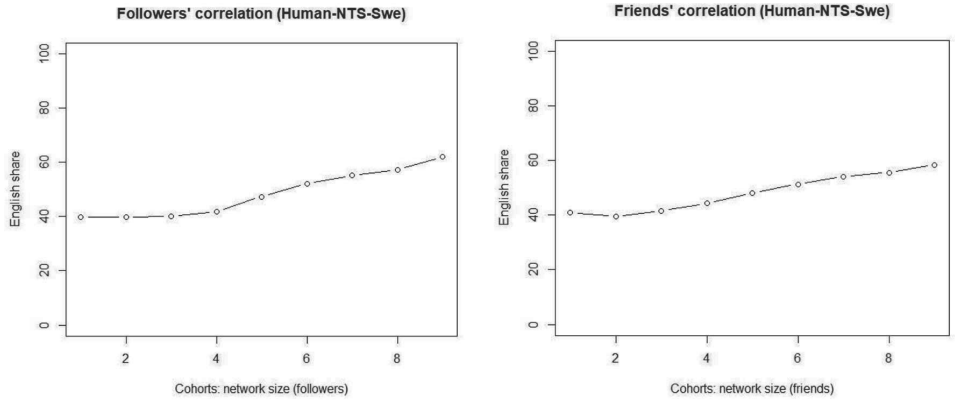


FIGURE 5 | Bot-free correlations of truly weak ties (**Left**) and slightly stronger ties (**Right**).

et al., 2011). Unfortunately, age is not included in the metadata parameters in the raw data, and its role cannot be controlled.

The main finding here is that we can confirm our pilot results in Laitinen et al. (2017). The new cohort-based findings using bot-free data suggest that network size plays a role in leveling the differences beyond a certain threshold. The following section will turn its attention to ego networks.

An Algorithmic Approach to Networks in Sociolinguistics

This section digs deeper into digital networks and uses an algorithmic method that complements the results above and provides tools for analyzing networks of mobile individuals. We operate with the 1-step neighborhood, which consists of a focal node, ego, and nodes directly connected to it. We also include the connections between nodes (degree 1.5). Twitter is a directed-graph network, and we are interested in what accounts "see" instead of how they are "seen," and consider friends rather than followers in the analysis. Consequently, we deal with a graph-based structure in which nodes represent accounts and directed edges are considered as a friend relationship, as in **Figure 6**, which visualizes two nodes in which A is either following B, or B is a friend of A.

The method assumes that account activities and mutual interaction between accounts have an impact on the relationship. To subject activities to the algorithms, we collected up to 3,200 recent tweets in JSON files for each account in the network and then extracted the values for how many times accounts in the entire network retweet or quote another account in the same network, and counted the number of times that accounts mention each other.

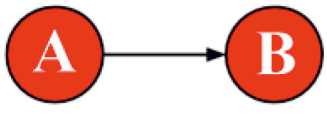


FIGURE 6 | A simplified example of a directed graph.

In order to extract ego-networks and to assess network values (either weak-tie and close-knit), we applied multiple criteria to the edges and nodes. While many of them are used in data mining, they measure network activities rather like the ethnographic methods in Milroy (1987) but applied to the parameters available in digital social networks.

First, we use a linear combination in (1), in which we assign weights to the links in the network.

$$\begin{aligned} \textit{Edge weight} \;&=\; (w_1 * \textit{retweet}_{count}) + \big(w_2 * \textit{quote}_{count}\big) \\ &+\; (w_1 * \textit{mention}_{count}) \end{aligned} \tag{1}$$

Where w_1, w_2, and w_3 are weights that can be assigned based on the application of interest so that $\sum_{i=1}^{3} w_i = 1$. Weights regulate the importance of each feature in the analysis. For instance, if we want to focus on the number of retweets, we assign $w\_1 = 1$ while w_2, $w\_3 = 0$. Moreover, we assume that those accounts that have a higher rate of publishing tweets have more impact on the information flow in a network, which should be considered as a factor. The point is to separate active accounts from those that use Twitter passively while rarely creating any content. To assign weights, we extracted the age (in days) of each account and the total number of tweets. Then, calculating the average number of tweets per day for each account and using (2), we can assign weights to the individual nodes as well.

$$\textit{Node weight}\;(A) = \frac{\textit{average tweets per day for account } A}{W}, \tag{2}$$

$$\textit{where}: \quad W = \sum_{i=1}^{N} \textit{average tweets per day for account } A_i. \tag{3}$$

Figure 7 visualizes an ego network with 30 nodes and 142 edges, (a) without assigning weights to the nodes and edges, and (b) by assigning weights using the formulae in (1)–(3). The larger the node, the higher the value for tweets per day, and the thicker the link, the stronger the connection between the nodes.

Second, we use *betweenness centrality* (BC) to detect the density and to interpret how people in a network are connected

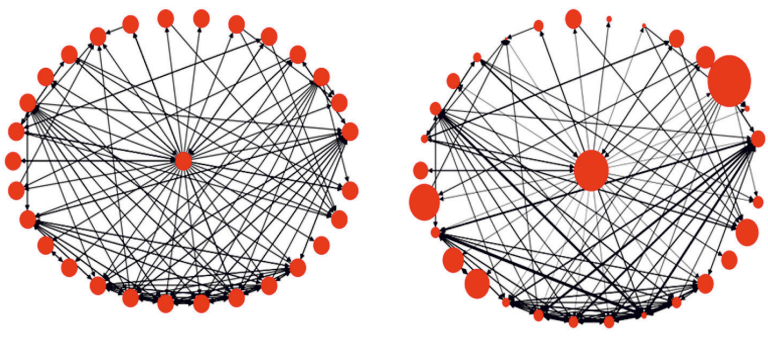


FIGURE 7 | An ego network, without assigning weights (**Left**), and with weights (**Right**).

to each other. The BC values represent the ratio with which an account establishes the shortest path between any pair in the network (Freeman, 1977). In other words, the BC of node v is the sum of the fraction of all of the shortest paths for any pair of nodes in the network that pass through v:

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \quad (4)$$

Where V is the set of nodes in the network, $\sigma(s,t)$ is the number of shortest paths between nodes s and t, and $\sigma(s,t|v)$ is the number of shortest paths between s and t that pass through V. Hence, the lower the BC value, the fewer the shortest paths passing through that account, and *vice versa*. The assumption is that the lower the spread (i.e., the difference between the higher and the lower values) of BC values in a network, the more connected the accounts are to each other, and the network consists of strong ties.

Consider **Figure 8**, in which the spread of the BC values is zero. The network is complete as all the nodes are connected to each other and the shortest path between each pair of the nodes is the direct path between those two nodes, and the path does not pass through any other nodes.

Figure 9 visualizes two real Twitter networks. The yellow nodes represent the ego, while the black links represent Two-way connections and blue links show One-way connections. Using visual cues, we can see that the left side is a weak-tie network, while the one on the right represents a stronger-tie network, and this is also supported by quantitative evidence. The spread value for the weak-tie network is 0.5455 and the corresponding value for the strong ties is 0.3014. We use normalized BC values to address the effect of network sizes on the calculations.

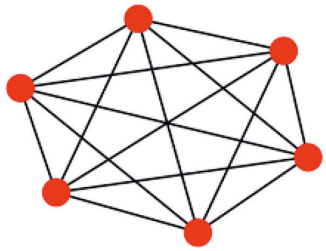


FIGURE 8 | A complete graph with 6 nodes and BC mean and spread.

The third measure is *closeness centrality* (CC), a concept that measures the distance between nodes (Perez and Germon, 2016). In the graph theory, the distance between two nodes is defined as the length of the shortest path between two nodes. CC is the reciprocal of the sum of the distances from a node to all the other nodes in the network. As in the case of the BC analysis, to eliminate the effect of network size we applied the normalized CC values in the analysis. The normalized CC value is calculated using the formula in (5):

$$C_C(v) = \frac{N-1}{\sum_{i=1}^{N-1} d(u,v)}. \quad (5)$$

Here, $d(u,v)$ is the shortest-path distance between u and v, and N is the number of nodes in the network. The CC values are between 0 and 1 for each node, and higher values of closeness on average could be interpreted as higher connection rates between nodes. In a directed graph in Twitter, there are two CC values for each node (i.e., incoming and outward). If the difference between the two CC values on average is low, it indicates that the majority of the connections in a network are Two-way links. Therefore, the network is a stronger-tie network.

The next two measures have been purpose-built by us and can be illustrated by inspecting the two networks in **Figure 9**, above. In the weak-tie network (left), the majority of the accounts are connected to each other through the ego node, while the accounts in the right-hand network are not only connected to the ego node but to the other accounts in the network as well, which means that the network consists of stronger ties. If we remove the ego node and its incoming/outgoing links from the data, we can then calculate the ratio of *disjoint nodes* in the network. We assume that the higher the value of the disjointness ratio, the weaker the network will be. Furthermore, as mentioned before concerning the edge weights, we can calculate the mean values of the edge

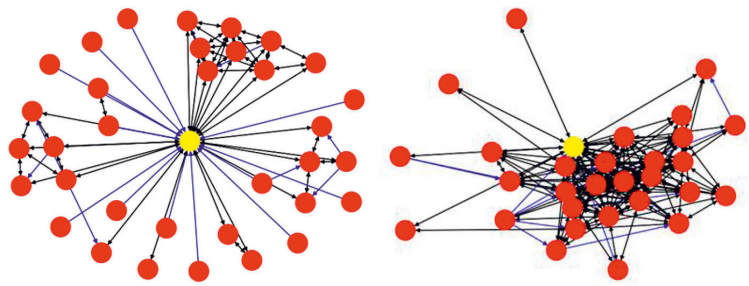


FIGURE 9 | A weak-tie ego network (**Left**) and a strong-tie network (**Right**).

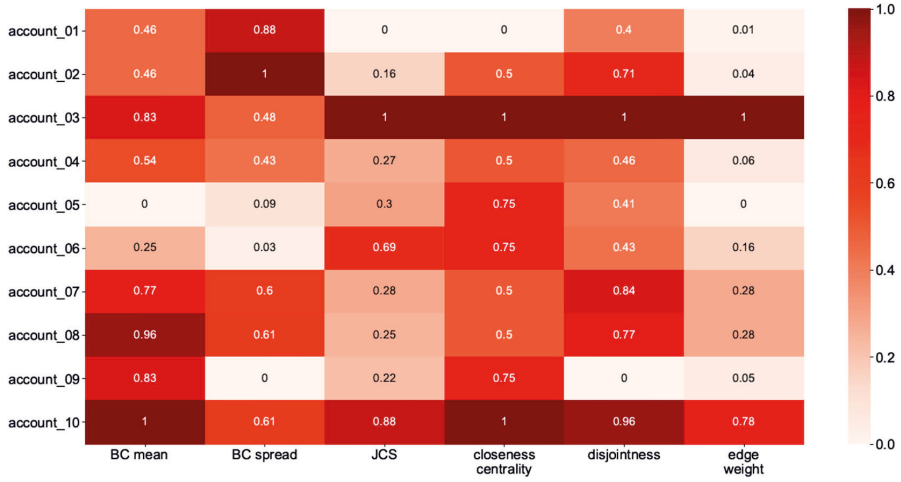


FIGURE 10 | Ten candidate accounts and their corresponding values for indices.

weights for each network. We would argue that, for a stronger-tie network, the mean value of the edge weights should be higher than the corresponding value for a weaker-tie network because individuals in a strong-tie network might be expected to have more interaction and activities than in a weaker-tie network.

The last algorithm strengthens the method by bringing in a tool that enables us to measure the similarity between two sets. It builds on the idea that individuals in a strong-tie network might be expected to be more similar to each other than individuals in a network characterized by weak ties. If we use Milroy's (1987) ethnographic work as our point of comparison, men in the Belfast neighborhoods were localized and spent more time with those who were similar to themselves in their dense strong-tie networks than women.

To measure similarity between sets, we use the Jaccard Similarity Coefficient (JSC). It is a symmetric measure that can be used to calculate the similarity between sets A and B as follows:

$$JSC = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

The assumption is that if two accounts have a high number of shared friends (i.e., a high JCS value), they are more similar to each other than two other accounts with a lower JCS value. Consequently, if the average JCS values for all the nodes in ego network A are higher than the averages for another network B, it means that the accounts in the A network are more similar to each other and that we are dealing with a stronger-tie network, and *vice versa*.

Consider the two networks presented in **Figure 9**, above. Using the formula presented in (6), we can calculate the mean JSC value for the weak-tie network to be 0.12 and the corresponding value for the stronger-tie to be 0.9. The average similarity for the network on the right is almost 8 times higher than the average similarity for the network on the left.

To measure the network qualities, we extracted the values for each network and, with the aid of Min-Max normalization, placed them on an interval [0,1]. We subtracted the calculated values for the BC mean, BC spread, disjointness ratio, and CC difference from 1 in order to make them comparable with the other features. The values are shown in **Figure 10**. The higher values for each feature (i.e., the darker the cell) indicate stronger-tie networks, and *vice versa*.

To assign labels (weak-tie or strong-tie) to the candidate networks, we calculated the mean values (strength coefficient *alpha*) for each cell in **Figure 10**. We then labeled the accounts with lower alpha values as weak-tie networks (W1–5) and the rest as strong-tie networks (S6–10), as shown in **Figure 11**.

The strength values (top) and the visualizations of all of the ten networks suggest that the algorithms are able to distinguish between networks with differing qualities. The visualization shows that the candidate networks as a whole can be roughly divided into weak-tie networks and networks with stronger ties. The method is robust and is not affected by smaller clusters that might appear, for instance, inside a weak-tie network. As a whole, therefore, we are able to suggest that the differences between the network types are supported by complex multidimensional quantitative data and visual cues. The next step is, then, to test to see whether differing network structures are reflected in the linguistic behavior.

In the last part of this study we investigate how the dependent variables, listed in section Materials and Methods above, are distributed among the network types. The accounts, their sizes, and the normalized frequencies (per 100,000 messages) of the

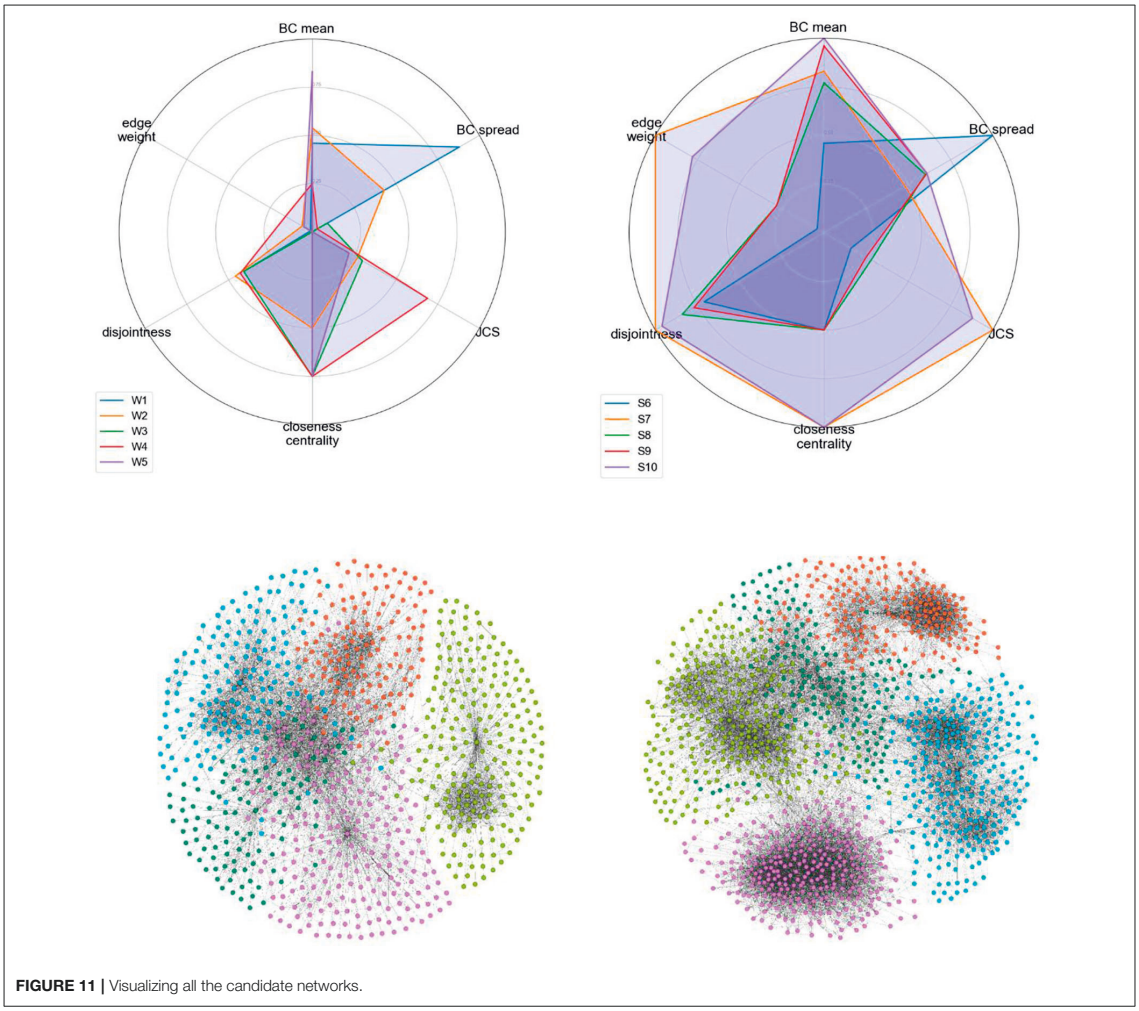


FIGURE 11 | Visualizing all the candidate networks.

dependent variables are shown in **Table 2**, below. The three columns on the right show the number of English messages in the network, the number of contractions in the text, and the frequency of *NEED to* + V constructions. It is important to note that, while the observations are based on a limited number of accounts, the data have been retrieved from the entire network connected to the ego node. These data consist of a total of 2,074 network nodes with over 2.6 million messages and nearly 30 million tokens of text. The network sizes vary, with the smallest possessing 105 nodes and the largest nearly 300. The number of messages varies between 100,915 and over half a million. The mean is 264,351 messages.

Figure 12 shows three boxplots that visualize the relationships between the weak- and strong-tie networks and the three

TABLE 2 | Statistics related to the dependent variables (normalized per 100,000).

Account	Network	N msg.	EngShare	Contr.	*NEED to* + V
W1	221	312,350	63,220	3,860	630
W2	175	150,774	40,910	940	310
W3	105	100,915	81,195	3,580	770
W4	132	192,944	45,237	3,230	560
W5	216	189,628	68,688	3,800	610
S6	166	253,758	79,534	2,590	840
S7	195	286,945	61,039	2,930	660
S8	281	316,944	85,387	5,790	890
S9	286	322,566	62,170	2,610	450
S10	297	516,686	81,261	6,290	1,070

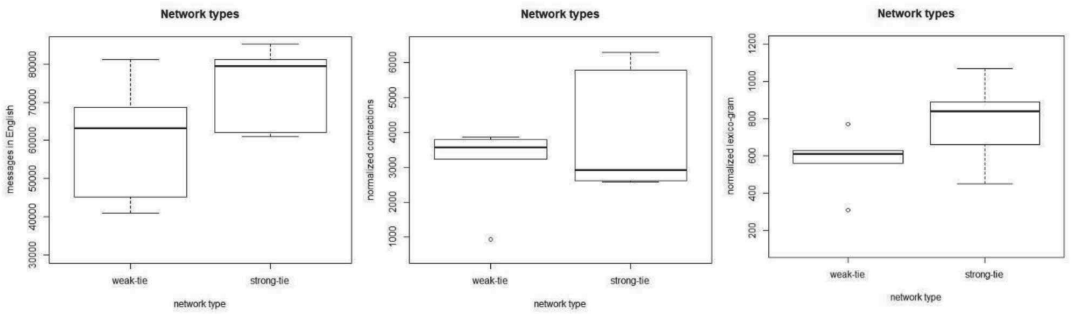


FIGURE 12 | The relationships between the network types and the dependent variables.

dependent variables. The data show no consistent pattern in which large networks would be quantitatively different from each other, but large weak and strong-tie networks behave similarly in terms of these variables. For the count of English messages (left), the mean value for the strong-tie networks is higher, but when tested with the Welch Two Sample *t*-test for independent samples, the differences between the networks are not statistically significant ($t = -1.55$, $p > 0.05$). The mean value for the contracted forms is slightly higher for the weak-tie networks, but the differences are not statistically significant ($t = -0.97$, $p > 0.05$). As for the lexico-grammatical variable, the mean is higher for the strong-tie networks, but the differences are not statistically significant ($t = -1.55$, $p > 0.05$).

The quantitative patterns observed are clear. When we investigate the large networks whose sizes are above the threshold level suggested in section A Cohort-Based Approach to Network Size, we can observe identical patterns. The results show no distinction between large weak-tie and strong-tie networks, which suggests that the differences observed in small ethnographic studies level out when the network size becomes sufficiently large. These observations support the cohort-based findings in section A Cohort-Based Approach to Network Size, above, and they also introduce ways of measuring the digital networks of mobile individuals in the social media.

We have attempted to demonstrate our algorithmic method which utilizes data-mining of the social media and uses a range of quantitative measures to establish network indices. The method enables us to establish networks of varying strengths and to determine that these varying qualities can not only be visually confirmed (**Figure 11**) but also supported by quantitative information. The method requires some computational power but still involves a qualitative element, since we have endeavored to ensure that the candidate networks represent similar content profiles. As we point out above, previous studies have suggested that various subpopulations have anomalously high network profiles (McCarty et al., 2001), and, at this stage, the objective has been to ensure that the candidate networks are similar. Our future objective is to test the algorithmic method with a far larger set of networks.

CONCLUSIONS

This article has investigated digital social networks of highly mobile individuals, and we have attempted to contribute to the study of social networks in sociolinguistics by providing tools for accessing large networks. The research objective has focused on the role played by network size as a key determinant in social networks. We have shown that network size has not been used in variationist sociolinguistics. Recent network studies in other fields have, however, suggested that network size could play an important role and that the distinction between network types might level out beyond a given threshold size of networks (Ma et al., 2019). Another of our motivations has been to observe real networks whose size is close to the average (at least in Western societies). The mean size of the ego networks (207 nodes) used in section An Algorithmic Approach to Networks in Sociolinguistics far exceeds the size of networks that have been covered in previous sociolinguistic studies, but they still fall within the limits of viable networks, as discussed in section Social Networks in Variationist Sociolinguistics.

As for the research questions, the first question focused on improving the methods used in sociolinguistics so that the quantitative variable of a network could be better operationalized in situations where the population consists of both socially and geographically highly mobile individuals. We have introduced two methods for accessing the networks of mobile individuals, thus expanding the empirical basis from small-scale ethnographic observations. Section A Cohort-Based Approach to Network Size introduced cohort-based methods, while in section An Algorithmic Approach to Networks in Sociolinguistics we detailed an algorithmic approach. The methods have a strong empirical basis, and they offer new tools for variationist sociolinguistics. They reveal fundamental differences in comparison with ethnographic approaches. For

instance, one of the advantages of ethnographic social network studies is that the methods build on the idea that networks are intrinsically a participant-related concept rather than something than an outsider analyst could construct (Milroy and Llamas, 2013). Our cohort-based method adopts an alternative approach, a clearly analyst-driven approach aimed at uncovering broad quantitative patterns in data rather than looking at existing networks. However, the algorithmic approach is very similar to the original idea, since the starting point is an existing network. As in Milroy and Milroy (1978) and Milroy (1987), the second method assumes the unit of study to be essentially a pre-existing category. Moreover, our method assumes network ties to be multidimensional, as the algorithms account for not only frequency of communication, but also a range of other factors. This means modernizing the network concept in sociolinguistics and bringing it closer to the contemporary idea that networks are not based on a simple dichotomy but consist of a range of attributes (Brashears and Quintane, 2018).

The second research question concentrates on the effect of network size on the validity of the theory by combining methods from sociolinguistics with computer science. Our results gained from both methods suggest that network size plays a role, and that the distinction between weak ties and stronger ties levels out once the network size grows beyond roughly 120 nodes. This finding is similar to the finding related to trust in networks (see section Social Networks in Variationist Sociolinguistics, above). We would, therefore, suggest that further studies be made of the digital networks of mobile individuals. Our raw data and the code are publicly available to other researchers.

Our future plans include continuing to work using the two methods. We plan to expand the cohort-based method and to test it with other dependent variables than simply language choice. Moreover, the metadata available in the tweet stream contain a number of possible predictors other than network size, and they need to be tested using linear regression. As for the algorithmic approach, our objective is to collect data from (tens of) thousands of accounts to scale up the method.

DATA AVAILABILITY STATEMENT

The Twitter dataset used in section A Cohort-Based Approach to Network Size is publicly available through the streaming API (https://developer.twitter.com). The data used in section An Algorithmic Approach to Networks in Sociolinguistics can be made available by the authors, without any undue restrictions, to qualified researchers. The code used in the algorithmic approach is available through GitHub (https://github.com/Masoud-Fatemi/Two-approaches-to-digital-social-networks).

AUTHOR CONTRIBUTIONS

This study was conceptualized by ML and MF. JL was responsible for data curation together with MF. The investigations were carried out by ML and MF. The methodology developed by MF, JL, and ML. The visualizations were created by MF and ML. The project was administered by ML, who was also responsible for writing the original draft version. All authors contributed to the article and approved the submitted version.

FUNDING

This research has been supported by the generous funding of the Center for Data Intensive Sciences and Applications (DISA) at Linnaeus University. In addition, ML has been able to invest research time offered by his position at the University of Eastern Finland.

ACKNOWLEDGMENTS

We wish to thank Professor Emerita Lesley Milroy for her valuable comments on this article. The usual disclaimers apply.

REFERENCES

Aral, S., and Van Alstyne, M. (2011). The diversity-bandwidth trade-off. *Am. J. Sociol.* 117, 90–171. doi: 10.1086/661238

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J. Mathem. Sociol.* 25, 163–177. doi: 10.1080/0022250X.2001.9990249

Brashears, M. E., and Quintane, E. (2018). The weakness of tie strength. *Soc. Networks* 55, 104–115. doi: 10.1016/j.socnet.2018.05.010

Chambers, J. K. (2003). *Sociolinguistic Theory. Linguistic Variation and its Social Significance. 2nd Edn*. Oxford: Blackwell.

Coats, S. (2017). "European language ecology and bilingualism with English on Twitter," in *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities*, eds C. Wigham and E. Stemle (Bozen/Bolzano: Eurac Research), 35–38.

Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *J. Hum. Evol.* 22:6, 469–493. doi: 10.1016/0047-2484(92)90081-J

Eleta, I., and Golbeck, J. (2014). Multilingual use of Twitter: Social networks at the language frontier. *Comp. Hum. Behav.* 41, 424–432. doi: 10.1016/j.chb.2014.05.005

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry* 40, 35–41. doi: 10.2307/3033543

Gonçalves, B., Loureiro-Porto, L., Ramasco, J., and Sánchez, D. (2018). Mapping the Americanization of English in space and time. *PLoS ONE* 13:e0197741. doi: 10.1371/journal.pone.0197741

Graedler, A-L. (2014). Attitudes towards English in Norway: a corpus-based study of attitudinal expressions in newspaper discourse. *Multilingua* 33, 291–312. doi: 10.1515/multi-2014-0014

Graham, M., Hale, S., and Gaffney, D. (2013). Where in the world are you? Geolocation and language identification in Twitter. *Profes. Geogr.* 66, 568–578. doi: 10.1080/00330124.2014.907699

Granovetter, M. (1973). The strength of weak ties. *Am. J. Sociol.* 78, 1360–1380. doi: 10.1086/225469

Granovetter, M. (1983). The strength of weak ties: a network theory revisited. *Sociol. Theory* 1, 201–233. doi: 10.2307/202051

Hale, S. (2014). "Global connectivity and multilinguals in the Twitter network," in *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems* (Toronto), 833–842.

Huang, Y., Guo, D., Kasakoff, A., and Grieve, J. (2016). Understanding US regional linguistic variation with Twitter data analysis. *Comp. Environ. Urban Syst.* 59, 244–255. doi: 10.1016/j.compenvurbsys.2015.12.003

Kim, S., Weber, I., Wei, L., and Oh, A. (2014). "Sociolinguistic analysis of Twitter in multilingual societies," in *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (Santiago), 243–248.

Krishnamurthy, B., Gill, P., and Arlitt, M. (2008). "A few chirps about Twitter," in *Proceedings of the first workshop on Online social networks (WOSN '08)*, eds. C. Faloutsos, T. Karagiannis and P. Rodriguez (New York, NY: ACM), 19–24. doi: 10.1145/1397735.1397741

Kuikka, V. (2018). Influence spreading model used to analyse social networks and detect sub-communities. *Comp. Soc. Netw.* 5:12. doi: 10.1186/s40649-018-0060-z

Labov, W. (2001). *Principles of Linguistic Change. Vol. 2. Social Factors.* Oxford: Blackwell.

Laitinen, M. (2016). "Ongoing changes in English modals: on the developments in ELF," in *New Approaches in English Linguistics: Building Bridges,* eds O. Timofeeva, S. Chevalier, A.-C. Gardner, and A. Honkapohja (Amsterdam: John Benjamins), 175–196.

Laitinen, M., and Lundberg, J. (2020). "ELF, language change and social networks: evidence from real-time social media data," in *Language Change: The Impact of English as a Lingua Franca,* eds A. Mauranen and S. Vetchinnikova (Cambridge: Cambridge University Press).

Laitinen, M., Lundberg, J., Levin, M., and Lakaw, A. (2017). "Revisiting weak ties: using present-day social media data in variationist studies," in *Exploring Future Paths for Historical Sociolinguistics,* eds T. Säily, M. Palander-Collin, A. Nurmi, and A. Auer (Amsterdam: John Benjamins), 303–325.

Laitinen, M., Lundberg, J., Levin, M., and Martins, R. (2018). "The Nordic Tweet Stream: a dynamic real-time monitor corpus of big and rich language data," in *Proceedings of Digital Humanities in the Nordic Countries 3rd Conference* (Helsinki). Available online at: CEUR-WS.org/Vol-2084/short10.pdf (accessed April 11, 2019).

Lamanna, F., Lenormand, M., Henar Salas-Olmedo, M., Romanillos, G., Gonçalves, B., and Ramasco, J. (2018). Immigrant community integration in world cities. *PLoS ONE* 13:e0191612. doi: 10.1371/journal.pone.0191612

Leech, G. (2013). "Where have all the modals gone? An essay on the declining frequency of core modal auxiliaries in recent standard English," in *English Modality: Core, Periphery and Evidentiality,* eds J. I. Marín-Arrese, M. Carretero, J. Arús Hita, and J. van der Auwera (Berlin: Mouton de Gruyter), 95–115.

Leppänen, S., Pitkänen-Huhta, A., Nikula, T., Kytölä, S., Törmäkangas, T., Nissinen, K., et al. (2011). *National Survey on the English Language in Finland: Uses, Meanings and Attitudes. (Studies in Variation, Contacts and Change in English, 5)*. Available online at: http://www.helsinki.fi/varieng/journal/volumes/05/ (accessed April 11, 2019).

Lippi-Green, R. (1989). Social network integration and language change in progress in a rural alpine village. *Lang.Soc.* 18, 213–234. doi: 10.1017/S0047404500013476

Lundberg, J., Nordqvist, J., and Laitinen, M. (2019). "Towards a language independent Twitter bot detector," in *Proceedings of the Digital Humanities in the Nordic Region (DHN2019)* (University of Copenhagen). Available online at: http://ceur-ws.org/Vol-2364/28_paper.pdf (accessed April 11, 2019).

Ma, X., Cheng, J., Iyer, S., and Naaman, M. (2019). "When do people trust their social groups?," in *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)* (New York, NY: ACM).

McCarty, C, Killworth, P., Bernard, R., Johnsen, E. C., and Shelley, G. A. (2001). Comparing two methods for estimating network size. *Hum. Organ* 60, 28–39. doi: 10.17730/humo.60.1.efx5t9gjtgmga73y

Milroy, J. (1992). *Linguistic Variation and Change*. Oxford: Blackwell.

Milroy, J., and Milroy, L. (1978). "Belfast: change and variation in an urban vernacular," in *Sociolinguistic Patterns in British English,* ed P. Trudgill (London: Edward Arnold), 19–36.

Milroy, J., and Milroy, L. (1985). Linguistic change, social network and speaker innovation. *J. Linguist* 21, 339–384. doi: 10.1017/S0022226700010306

Milroy, L. (1987). *Language Change and Social Networks. 2nd Edn.* Oxford: Blackwell.

Milroy, L., and Llamas, C. (2013). "Social networks," in *The Handbook of Language Variation and Change,* eds J. K. Chambers and N. Schilling (Oxford: Blackwell), 407–427.

Milroy, L., and Milroy, J. (1992). Social network and social class: toward an integrated sociolinguistic model. *Lang. Soc.* 21, 1–26. doi: 10.1017/S0047404500015013

Modiano, M. (2003). Euro-English: a Swedish perspective. *Eng. Today* 19, 35–41. doi: 10.1017/S0266078403002074

Nevalainen, T. (2000). Mobility, social networks and language change in early modern England. *Eur. J. Eng. Stud.* 4, 253–264. doi: 10.1076/1382-5577(200012)4:3;1-S;FT253

Nguyen, D., Dogruöz, S., Rosé, C., and de Jong, F. (2016). Computational sociolinguistics: a survey. *Comput. Ling.* 42, 537–593. doi: 10.1162/COLI_a_00258

Pan, W., Ghoshal, G., Krumme, C., Cebrian, M., and Pentland, A. (2013). Urban characteristics attributable to density-driven tie formation. *Nat. Commun.* 4:1961. doi: 10.1038/ncomms2961

Perez, C., and Germon, R. (2016). Graph creation and analysis for linking actors: application to social data. *Autom. Open Source Intell.* 7, 103–129. doi: 10.1016/B978-0-12-802916-9.00007-5

Preisler, B. (2003). English in Danish and the Danes' English. *Int. J. Soc. Lang.* 159, 109–126. doi: 10.1515/ijsl.2003.001

Raumolin-Brunberg, H. (1996). "Social factors and pronominal change in the seventeenth-century: The Civil War effect?" in *Advances in English Historical Linguistics,* eds J. Fisiak and M. Krygier (Berlin: Mouton de Gruyter), 361–388.

Tagliamonte, S. A., and D'Arcy, A. (2009). Peaks beyond phonology: Adolescence, incrementation, and language change. *Language* 85, 58–108. doi: 10.1353/lan.0.0084

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Paper **II**

Fatemi, M., Kucher, K., Laitinen, M., & Fränti, P. (2021)
"Self-similarity of Twitter users"
2021 Swedish Workshop on Data Science, 1–7
https://doi.org/10.1109/SweDS53855.2021.9638288

Self-Similarity of Twitter Users

Conference Proceedings: 2021 Swedish Workshop on Data Science (SweDS)
Author: Masoud Fatemi
Publisher: IEEE
Date: 02 December 2021

Quick Price Estimate

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.htmtb learn how to obtain a License from RightsLink.

If applicable, University MicroFilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

I would like to... reuse in a thesis/dissertation

Self-Similarity of Twitter Users

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

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

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

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

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

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

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

I. Introduction

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

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

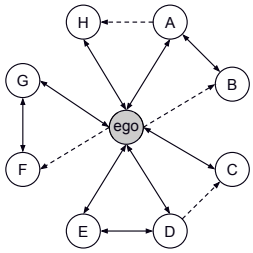
As discussed in Section I-C, we adopt the perceived-similarity approach. We ask Twitter users to provide a list of accounts that are similar to themselves and then compare this list to large user-generated data of interactions. The objective is to identify which interaction category most effectively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

A. Twitter Ego Networks

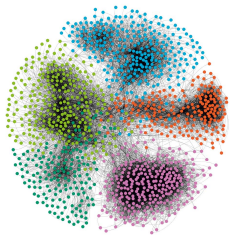
A literature review from computer science and social anthropology reveals that *ego networks* are the cornerstones in studying social networks [9], [10]. They are the primary structural characteristics of individual networks [9]. Indeed, the concept of an ego network is essential when identifying key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

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

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



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



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

Fig. 1. Two examples of social networks.

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

B. User Similarity on Twitter

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

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

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

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

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

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

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

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

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

C. Data

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

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

TABLE I
TWITTER DATA STATISTICS

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

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

II. DETECTING SIMILAR USERS IN SOCIAL NETWORK

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

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

A. Interaction-Based Similarity

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

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

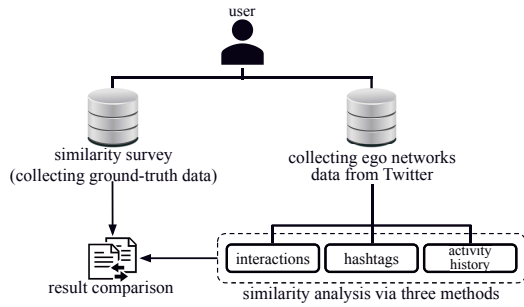


Fig. 2. Overview of the proposed model.

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

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

B. Profile-Based Similarity

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

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

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

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

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

C. Hashtag-Based Similarity

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

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

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

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

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

III. Empirical Results

A. Detecting Similarity of Twitter Users

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

TABLE II
ACTIVITY-BASED FEATURES

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

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

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

(a)

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

(b)

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

TABLE III
A: JOHN'S HASHTAGS

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

TABLE IV
B: CANDI'S HASHTAGS

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

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

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

B. Effect of Network Size

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

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

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

Fig. 4. Example results of similar accounts extraction.

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

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

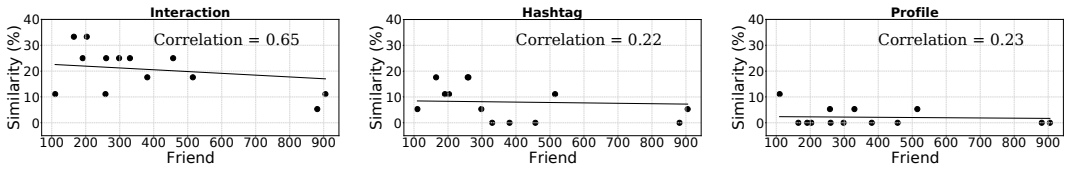


Fig. 5. Correlation analysis between the size of ego networks (number of friends) and the accuracy of three proposed approaches for measuring user similarity. The correlation values of the network sizes and similarity categories (interaction, hashtags, and activity) are -0.65, -0.22, and -0.23, respectively.

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

C. Male Ego Networks vs. Female Ego Networks

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

IV. Conclusion

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

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

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

Acknowledgment

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

References

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

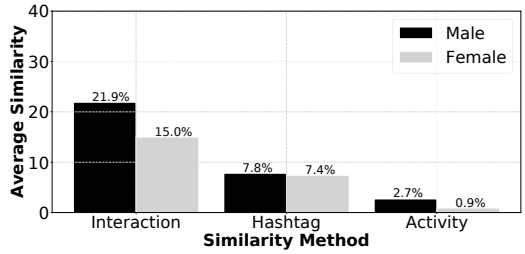


Fig. 6. A comparison of male vs. female accounts w.r.t similarity accuracy.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Paper **III**

Fatemi, M., Sieranoja, S., Laitinen, M., & Fränti, P. (2025)
"Detecting connectivity patterns in Nordic Twittersphere by cluster analysis"
SN Computer Science, 6(7), 815
https://doi.org/10.1007/s42979-025-04353-y

SPRINGER NATURE

Detecting Connectivity Patterns in Nordic Twittersphere by Cluster Analysis

Author: Masoud Fatemi et al
Publication: SN Computer Science
Publisher: Springer Nature
Date: Sep 10, 2025

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.

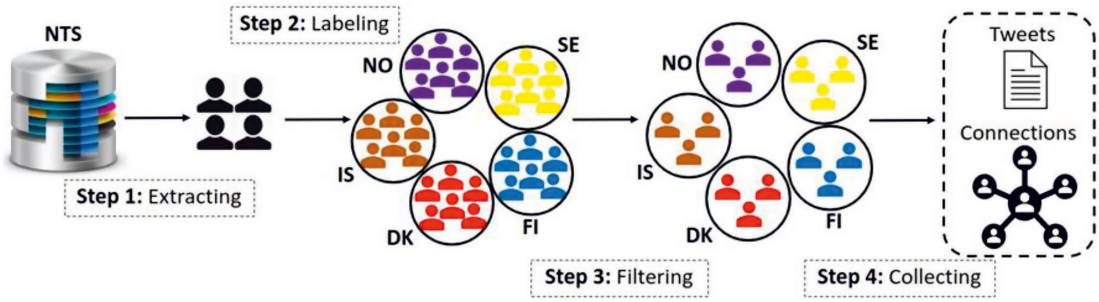


Fig. 1 An overview of NTN construction steps

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

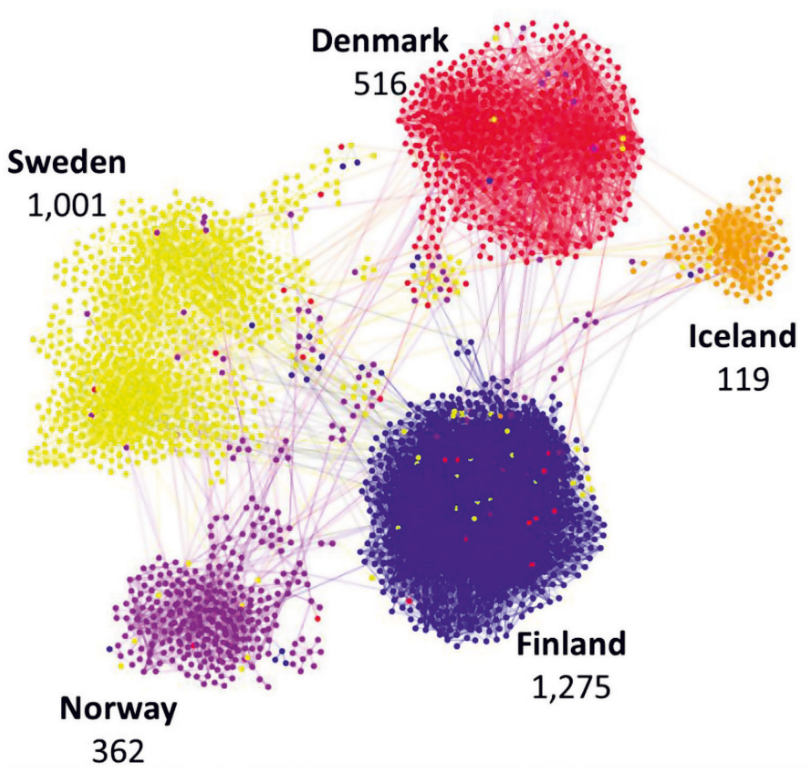


Fig. 2 A sample of *NTN-2022* dataset visualized according to the country of the users

[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

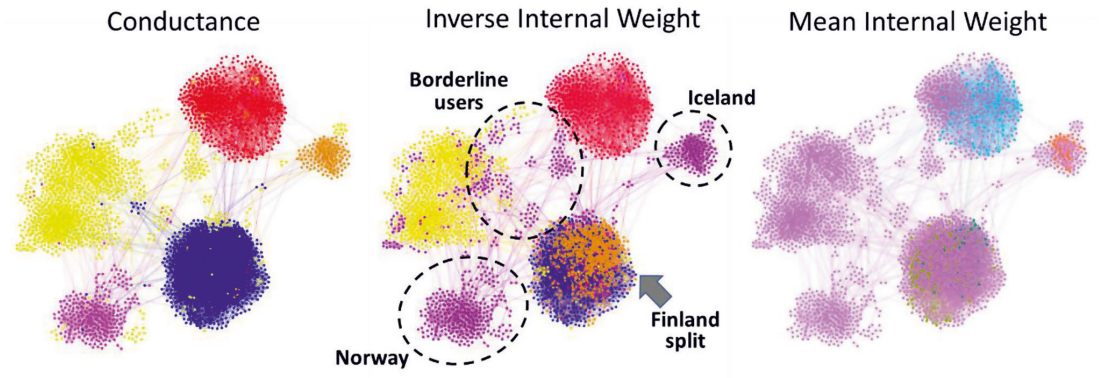


Fig. 3 Clustering with three different objective functions

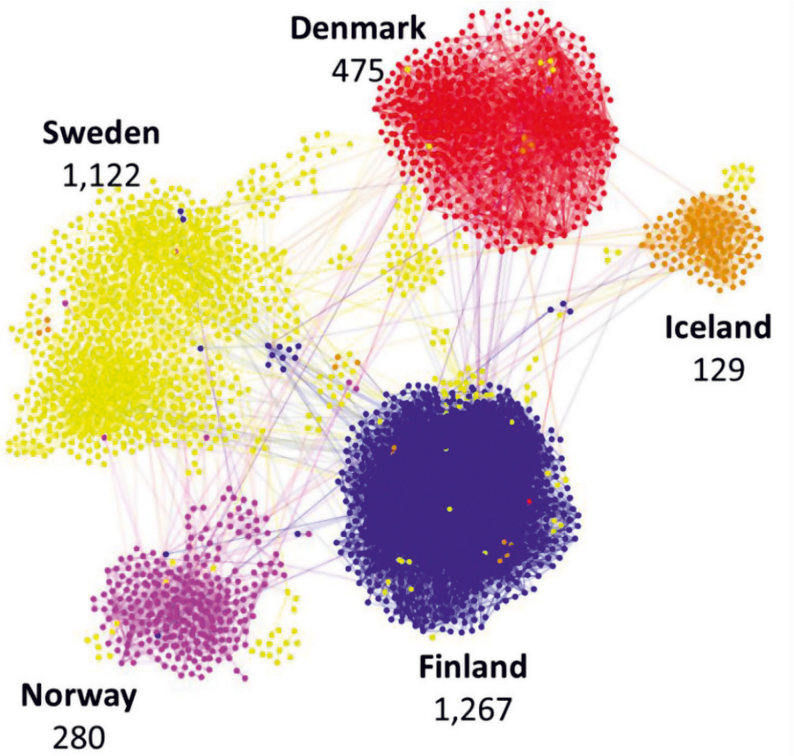


Fig. 4 Detected five clusters using the M-algorithm with conductance function. The result matches very closely the home countries of the users

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.

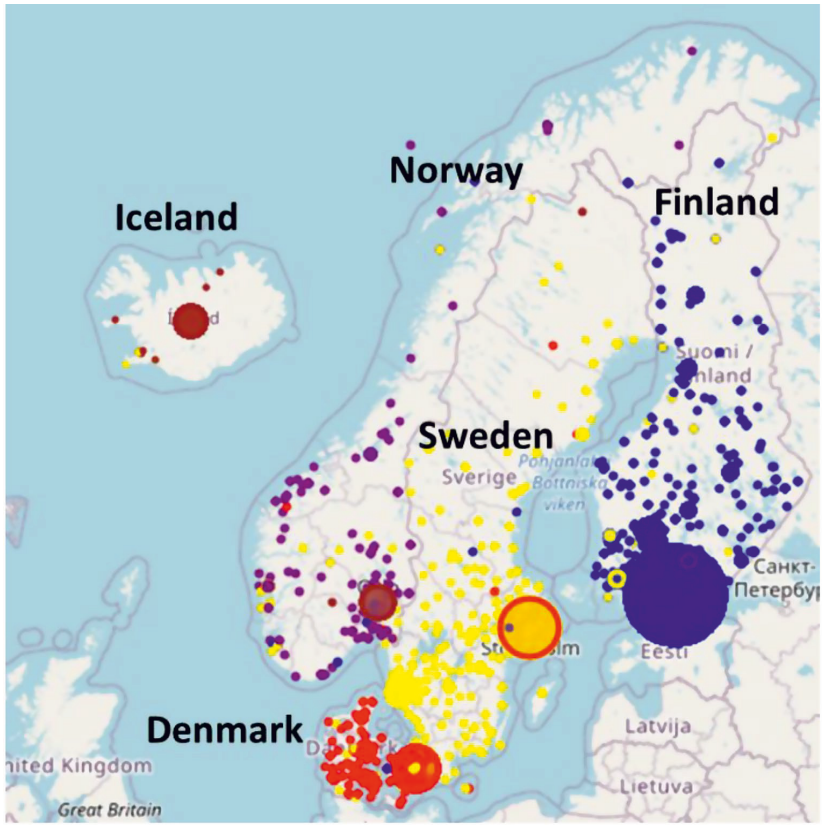


Fig. 5 Clustering results on map. Users are plotted at their home location and colored by the cluster it belongs to (Finland = blue, Sweden = yellow, Norway = purple, Denmark = red, Iceland = brown). Note: The geographical location of Iceland is artificially moved closer to Norway for making the figure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

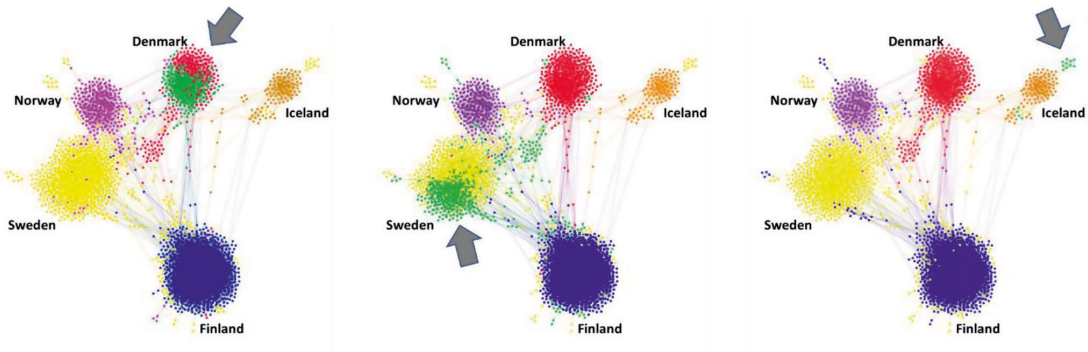


Fig. 6 Having six clusters does not lead to stable clustering results, and the additional cluster is highly sensitive to the initialization

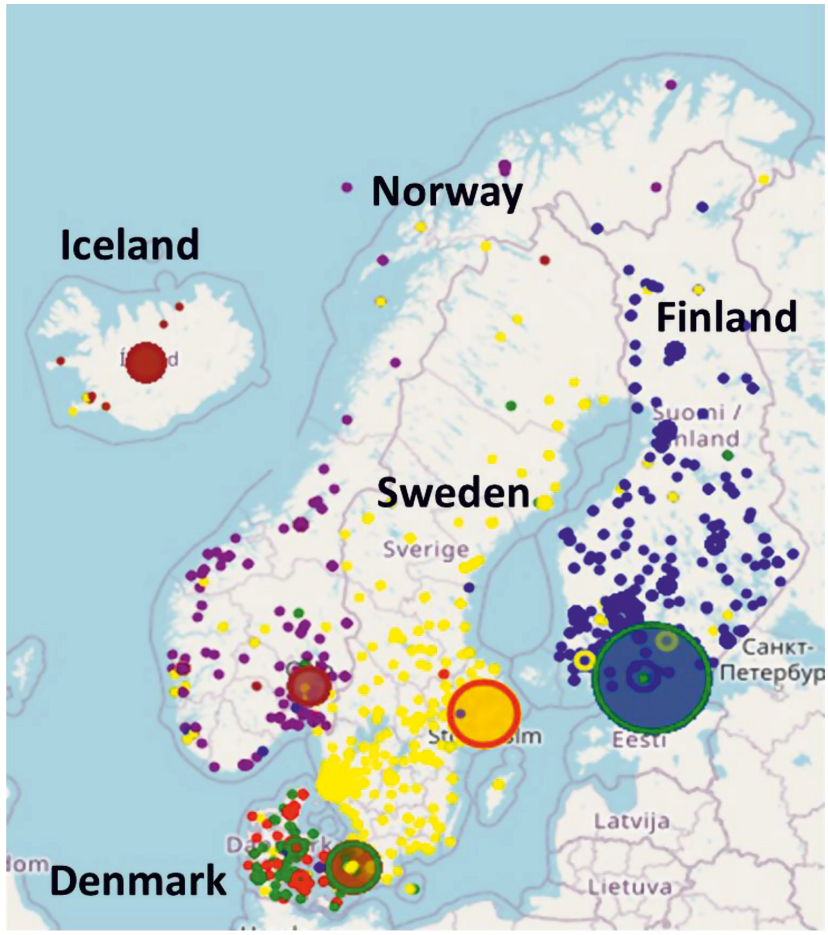


Fig. 7 Geographical distribution of the clusters in case when the additional cluster is allocated to Denmark. The statistics also showed high value of between these two clusters (14%) compared to the corresponding value of the Denmark cluster (5.2%)

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.

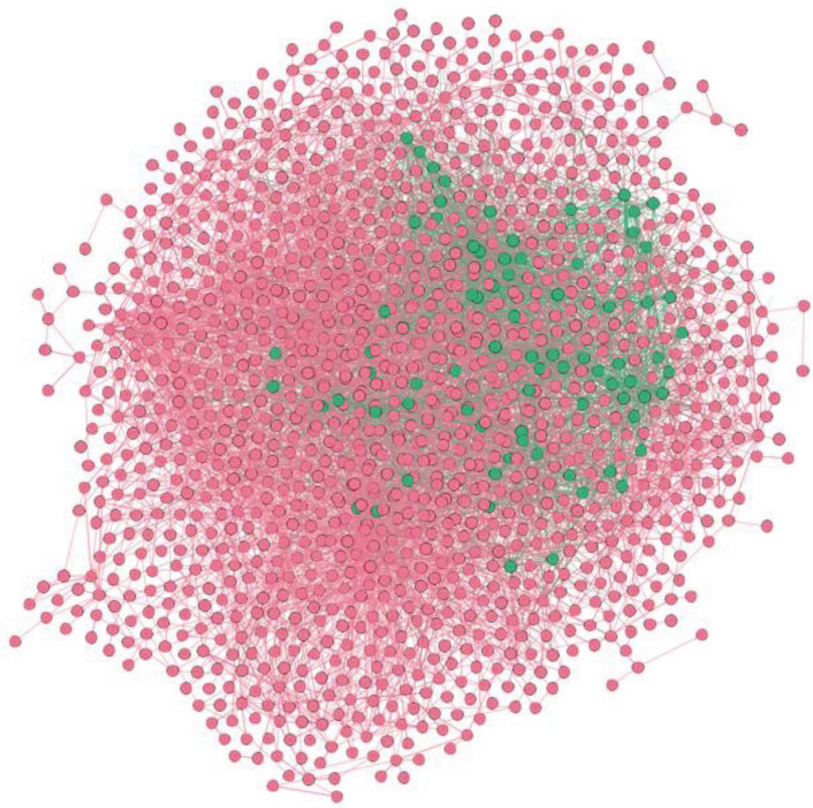


Fig. 8 Dividing Finland to two clusters seemed arbitrarily chosen without any natural explaining factor. The statistics also showed same value of between these two clusters (1%) compared to the corresponding value of the Finland cluster (1%)

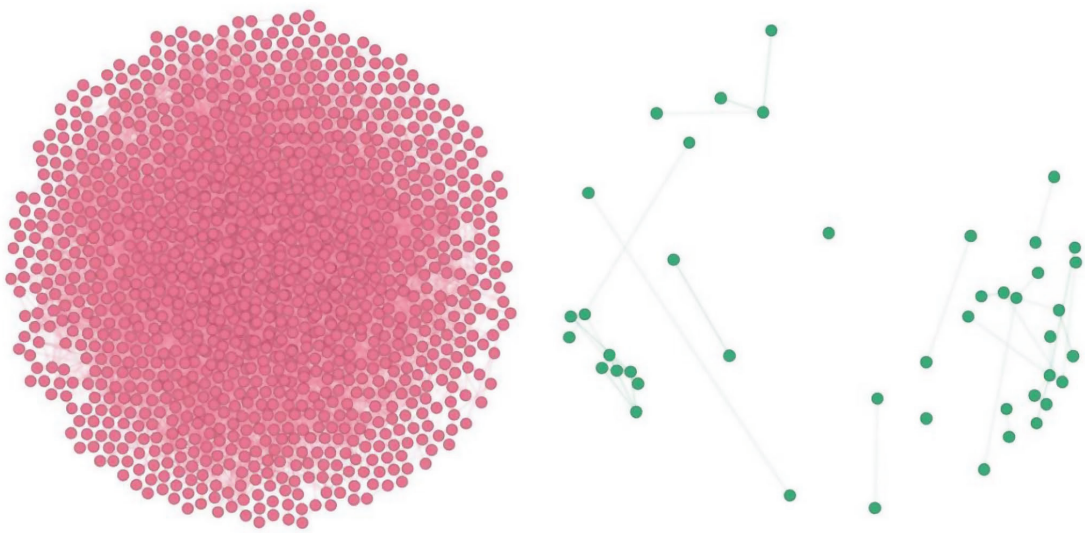


Fig. 9 The two sub-clusters inside Finland cluster consist mainly one large cluster. The second smaller cluster consists only isolated and weakly connected sub-clusters, i.e., outliers

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/S02660784030107X.

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.

Paper **IV**

Fatemi, M., Laitinen, M., & Fränti, P. (2026)
"Computer-mediated communication and networks: Quantifying ego network strength"
In Special Issue on Computer-Mediated Communication Corpora,
Language@Internet

The paper is accepted for publication.
The following is the author's latest version.

Paper **V**

Fatemi, M., Laitinen, M., & Fränti, P. (2025)
"Clustering digital ego networks by tie strength:
A scalable, platform-independent method"
The 20th International Conference on Intelligent Systems and Knowledge Engineering
Shunde, China The paper received the best paper award.

MASOUD FATEMI

Analysing online social networks, as the primary arenas of social interaction, is challenging and complex at scale and requires approaches beyond traditional methods. This thesis develops computational methods for modeling and clustering large-scale Twitter networks, focusing on tie strength, user similarity, and community structure. Combining multi-dimensional measures enables scalable, interpretable analysis while exposing the limits of existing methods for capturing complex social behavior.

UNIVERSITY OF
EASTERN FINLAND

uef.fi

PUBLICATIONS OF
THE UNIVERSITY OF EASTERN FINLAND
Dissertations in Science, Forestry and Technology

ISBN 978-952-61-6012-2
ISSN 2954-131X