Connectivity and community structure of online and register-based social networks

Menyhért, M.; Bokányi, E.; Corten, R.; Heemskerk, E.M.; Kazmina, Y.; Takes, F.W.

DOI
10.1140/epjds/s13688-025-00522-4

Publication date
2025

Document Version
Final published version

Published in
EPJ Data Science

Menyhért et al. *EPJ Data Science* (2025) 14:8
https://doi.org/10.1140/epjds/s13688-025-00522-4

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

Connectivity and community structure of online and register-based social networks

Márton Menyhért[1†], Eszter Bokányi[2*†], Rense Corten[3], Eelke M. Heemskerk[2], Yuliia Kazmina[2] and Frank W. Takes[1]

*Correspondence: e.bokanyi@uva.nl
[2]University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
Full list of author information is available at the end of the article
[†]Equal contributors

Abstract

The dominance of online social media data as a source for large-scale social network studies has recently been challenged by networks constructed from state-curated register data. In this paper focused on the cross-comparison of the network structures, we investigate the similarities and differences of the Dutch online social network (OSN) Hyves and a register-based social network (RSN) of the Netherlands. First and foremost, we find that node metrics and the connectivity of the two population-scale networks are similar, with more long-distance ties captured by the OSN, and with the OSN ties proving to be predictive of RSN ties. These results hold when correcting for population size and geographical distance, notwithstanding that these two factors appear to be the main drivers of connectivity. Second, we show using multiple algorithms that the community structure of the two networks is similar and that neither follows strict administrative geographical delineations (e.g., provinces). Instead, communities appear to either center around large metropolitan areas or, outside of the country's most urbanized area, comprise large blocks of interdependent municipalities. Beyond population and distance-related patterns, communities also highlight the persistence of deeply rooted sociocultural communities such as the Dutch Bible belt. The findings presented in this work aid in interpreting results from future studies in which register-based social networks are used to obtain insights into the social network structure of an entire population.

Keywords: Network science; Online social network; Register-based network; Spatial networks; Social connectedness; Community detection; Cross-comparison

1 Introduction

The recent rise in large-scale digital tracing from mobile devices enabled, for the first time, the study of large populations at once. These digital data sources offer researchers unprecedented access to huge amounts of information on human interactions and behavior (Lazer et al. [32]), enabling large-scale network analysis using data from online social networks (OSN) or mobile phone communication records (Kumar et al. [27], Onnela et al. [43], Hidalgo and Rodriguez-Sickert [20], Lambiotte et al. [29], Eagle et al. [13]). Early studies uncovered structural properties of global OSNs such as Meta (formerly Facebook) or X (formerly Twitter) (Ugander et al. [54], Myers et al. [40]), as well as more localized ones

Springer

such as Hyves in the Netherlands (Corten [9]) or iWiW in Hungary (Lengyel et al. [34]). Subsequent work illustrated the importance of large-scale social network structure by relating it to various socio-economic outcomes: the abundance and diversity of connections in a region is linked to its economic prosperity (Eagle et al. [12]), socio-economic inequality is reflected in more fragmented or closed network structures (Tóth et al. [53], Kovács et al. [25]), and having more people of high socioeconomic status in one's social network is associated with upward social mobility (Chetty et al. [7], Chetty et al. [8]). However, a drawback of these data sources is that it is often difficult to obtain reliable demographic information on the nodes, edges may represent multiple different types of connections, and the sample of both the nodes and the edges can be biased (Sloan et al. [50], Kassa et al. [22], Mislove et al. [39]).

Recently, state-curated registers that enable social network analysis for entire populations have become a promising alternative to OSNs and mobile phone data (Van der Laan et al. [56], Cremers et al. [10]). Register-based social networks (RSN) capture social connections reflected in administrative registers for a well-delineated population of a country (van der Laan and de Jonge [57], Cremers et al. [10]). Such a population-scale approach ensures high-quality data on nodes and edges with an administratively documented definition of both. As an example, Statistics Netherlands combined multiple registers covering family, school, work, household, and next-door neighbor connections into a comprehensive population-scale social network with multiple edge types (van der Laan [55]). A particular strength of RSN data is that it contains explicit information about the social foci where edges are created (school, work, family, etc.) and therefore allows for multilayer social network analysis. As such, the RSN describes the so-called social opportunity structure (Bokányi et al. [4], Kazmina et al. [23]), i.e., a variety of social contexts where interactions are possible, although it remains uncertain whether these interactions are actually realized. Despite its drawbacks such as an overrepresentation of social ties which are not necessarily active, and an underrepresentation of informal or past connections, this novel data source has already proven instrumental in offering new insights into persistent social issues such as segregation (Kazmina et al. [23]) or the intergroup connectivity of migrants and natives (Soler et al. [51]).

Neither OSNs nor RSNs were originally designed for research, resulting in different opportunities and challenges for each. Crucially, they are based on different conceptualizations of what constitutes a social tie. The ties in RSN data represent administratively defined relationships — people are connected through *formal ties* of kinship, work and school affiliations, and their registered address. OSNs, however, connect people through *informal ties* of nominating others as 'friends' or 'contacts'. This may include a myriad of social ties, such as family and friends, but also aspired relations, such as ties to celebrities. The findings of Van Eijk [58] suggest that they may compare well, as a large share of informal ties come from current or former formal ties of people, such as work, school, or family connections. In previous literature comparing online and offline networks, the offline part was not large-scale; nevertheless, the Dunbar number as limit for the number of close social relationships has been showed in Facebook and Twitter data ([11]), online connections were shown to reflect face-to-face encounters (Subrahmanyam et al. [52]), and students' online and offline connection structure might differ based on their cultural backgrounds (Cardon et al. [6]). However, it is not obvious whether the factors that influence connection probability play out similarly for large-scale offline formal and online

informal ties, such as homophily (McPherson et al. [38]), triadic closure (Asikainen et al. [1]), or geographic distance between people (Lambiotte et al. [29], Liben-Nowell et al. [35]).

We aim to position this work as a network structure comparison study. The focus is on cross-validating two large-scale, fundamentally different social network types (an online social network and a register-based social network) and extracting insights about their similarities and differences. This is an essential step in understanding the extent to which findings from the newly emerging register-based networks generalize to other types of social networks. We focus our comparison on connectivity and community structure to understand whether OSNs and RSNs exhibit the same topological features, since they represent different relationships. We ask a number of important questions related to how the OSN and RSN compare. Are connection patterns and structural features of nodes similar in the two networks? What are the types of connections (e.g. family, work, or school) from RSNs that are best represented in OSNs? How do population size and geographical distance, i.e., factors known to affect social network connections (Liben-Nowell et al. [35], Lambiotte et al. [29], Krings et al. [26], Xu et al. [60]), impact each of the two networks? Are meso-scale network structures, such as communities (Fortunato [16], Expert et al. [15], Fortunato and Hric [17]), persistent across different data sources? Are these communities indicative of predefined administrative delineations, or do they reveal other patterns of group connectivity in the population-scale social network?

We address the above questions using the unique combination of the Dutch OSN Hyves of 6.2M nodes and 320M edges, and a register-based multilayer social network (RSN) of the entire population of the Netherlands of 16.6M nodes and 570M edges, aggregated at the municipality level. We find that the municipality-level OSN and RSN compare remarkably well: the weighted networks are both similar for structural node metrics such as degree, clustering coefficient, and pagerank, and in the edge weights. When we remove the effects of population and distance as the main drivers of connectivity, these similarities still hold. Multivariate regression confirms that OSN edge weights are both significant and increase predictive power in modeling the RSN edge weights. The community structure of the two networks is also similar, and neither follows strict administrative geographical delineations but instead reveals deeply rooted sociocultural patterns.

In general, the findings presented in this paper confirm that both online social networks and register-based networks are useful for modeling the social network structure of a population. Moreover, both approaches lead to similar conclusions, since node-level topology, connectivity and community structure are similar in the two networks, especially at shorter distances, i.e. local spatial scales. As across different countries, more and more population-scale register-based social networks are expected to be available for research in the future (Magnani et al. [36], Savcisens et al. [48]), our paper gives a guideline on the cross-comparison and transferability of results between these different types of datasets.

2 Results

In this section, we first give a brief overview of the OSN and RSN datasets we use in this study. Then we provide two sets of comparative empirical results. The first pertains node properties and *connectivity*, and compares the weighted degree, the clustering coefficient, and the pagerank of nodes, and the number of edges between municipalities in both networks. We also investigate how the different layers in the RSN compare to the OSN. The

second set of experiments dives into the *community structure*, focusing on a comparison with administrative borders and sociocultural aspects of the connectivity patterns. In both sets of experiments, we consider normalization of the connection strength by two common aspects known to influence connectivity: population size and geographical distance.

2.1 Online and register-based population-scale social network datasets

We use a unique combination of two large-scale networks, each from 2010. The first is the formerly highly popular Dutch online social network Hyves (Corten [9]). It was the main online social networking site in the Netherlands before the advent of Facebook. The Hyves dataset is an anonymized version of the service provider used in the study of Corten [9] and includes 6.2M users and more than 320M edges. The second is a register-based social network sourced from Statistics Netherlands (van der Laan [55]). It contains all 16.6M registered residents of The Netherlands as of 1 January 2010. The roughly 570M edges between the nodes include formal ties of current family, school, and work relationships sourced from administrative databases. Family relations are derived from the official parent-child and partner registers, school relations are projections of the bipartite person-school network sourced from educational records, and work relations are sampled based on the bipartite employer-employee matching. These different types of edges constitute the layers of the RSN; we refer to all three of them as the combined layers. The data was accessed and analyzed using the Remote Access environment of Statistics Netherlands. The works of Corten [9] and Bokányi et al. [4] contain more details on the composition and topological properties of both networks.

Direct person-level matching of the two networks is legally and technically impossible due to privacy concerns and limited information available on the nodes. To facilitate the comparison, we aggregated the two networks to the level of 431 municipalities of the Netherlands as of 2010. Weighted edges between the municipality pairs denote the number of ties between people in the given network – the family, work, and school relations are all counted as one connection in this aggregation. Figure 1A and 1B illustrate the 500 edges with the largest weights in both networks. These strongest edges connect both major cities in the urban areas in the west but also reach out to the rural areas in the east and north. For more details on the two networks and the construction of the municipality-level aggregation, see the Data and methods section.

2.2 Connectivity in the OSN and the RSN

In the following set of empirical analyses, we seek to understand the structure of the OSN and the RSN in terms of their social connectivity. We do so for three versions of the network: the weighted network constructed from the aggregation of the person-level network reflecting the number of connections between municipalities, and for the same network with edge weights first normalized by population and then by population and geographical distance (see Data and methods for details). For each of the three variants of the network, we then look at structural node metrics of weighted degree, weighted clustering coefficient and pagerank, and the comparison of edge weights.

Frist, we compare the edge weight representing the connection counts between the municipality pairs in the OSN and in the RSN in Fig. 1C. We calulcate the Pearson correlation of the logarithm of the weights to decrease the influence of large outliers coming from the skewed weight distributions (see Additional file 1 Fig. 5 for all edge weight distributions)

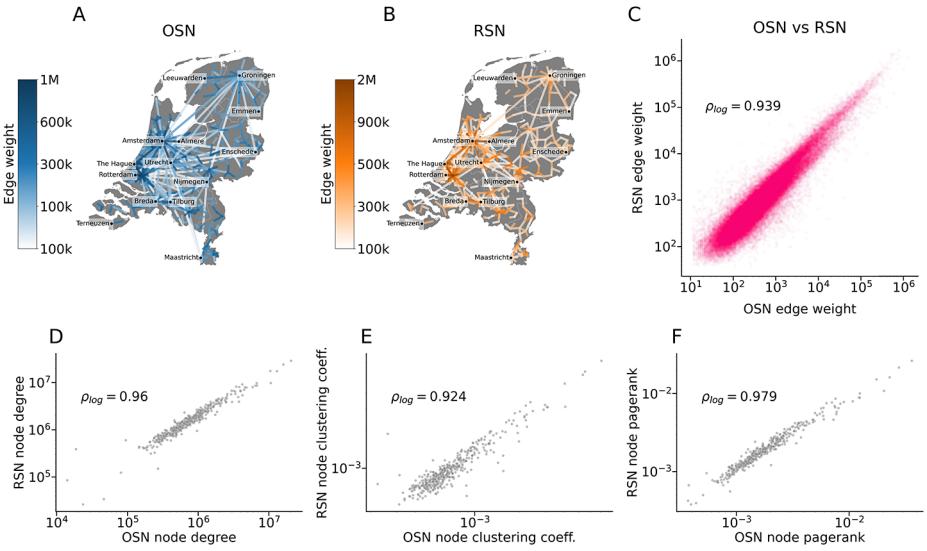

Figure 1 *The OSN and the RSN and the comparison of edge weights and node-level metrics.* Spatial layout of the top 500 edges between municipalities in *(A)* the OSN and *(B)* RSN. *(C)* Scatterplot of OSN and RSN edge weights between municipality pairs. *(D-F)* Scatterplot of OSN and RSN node properties: weighted degree, weighted clustering coefficient, and pagerank values

– which is 0.939. This suggests a high similarity between the two municipality-level networks. However, this similarity could be driven by factors that are known to influence the number of connections, such as population size and geographical distance between municipalities (Expert et al. [15], Krings et al. [26], Xu et al. [60], Lengyel et al. [34], Bokányi et al. [5], Lee et al. [33], Illenberger et al. [21]). We therefore normalize by population and distance to exclude these effects. Furthermore, we do a pairwise comparison between the OSN and each of the separate layers in the RSN (e.g., family, school, or work).

Second, Fig. 1D shows the scatterplots of weighted degree, weighted clustering coefficient, and pagerank in both the municipality-level OSN and the RSN, where edge weights are the connection counts. The Pearson correlation of the logarithm of the values – which we again use because of the skewed underlying distributions of these metrics (see Additional file 1 Figs. 2-4 for the distributions) is 0.960 for degree, 0.924 for clustering, and 0.979 for pagerank, indicating a high structural similarity between the nodes.

Figure 2A shows the correlation of edge weights between the OSN and the RSN (Combined), as well as for the family, school, and work layers separately. We calculate these correlations for three different edge weighting strategies. The Count captures the number of connections between two municipalities which we already used in Fig. 1. For each barplot from Fig. 2A, the first row shows the correlation corresponding to this weighting strategy for the differently composed RSN networks. The first bar on top shows the very high Pearson correlation of 0.939 from Fig. 1C. If we restrict the RSN to a single layer type in the subsequent bar plots, we get slightly lower correlation values: 0.924 for the family, 0.865 for the school, and 0.866 for the work layer. This comparison shows that for the count edge weights, the full RSN compares best with the OSN.

The population size of a municipality is an important driving factor for connectivity, as larger municipalities naturally have more connections. The Social Connectedness Index (SCI) metric of Bailey et al. [2] corrects for this dependence by normalizing the edge

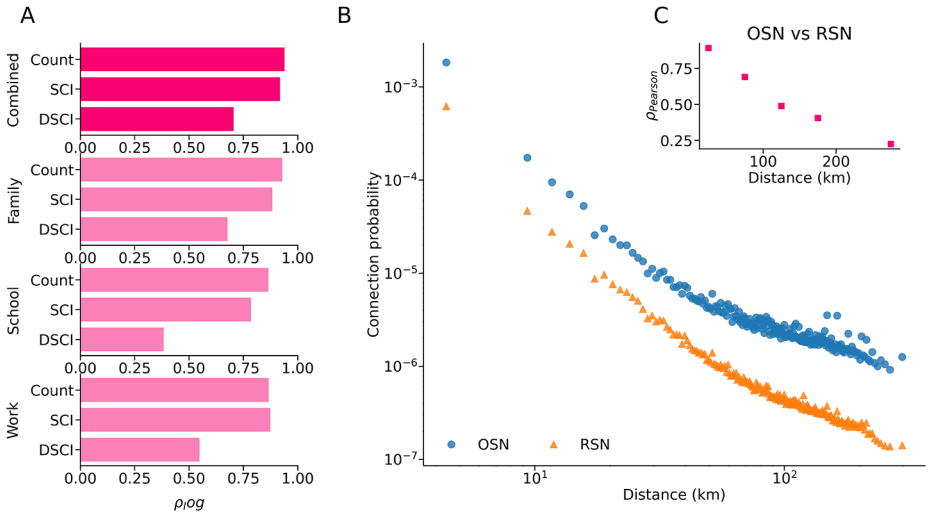

Figure 2 *Detailed correlation of edge weights and distance behaviour of connection probability of the OSN and the RSN. (A)* Pearson correlation similarity of OSN and RSN edge weights using all RSN layers (Combined), and Family, Work, and School layer separately, each for different edge weighting strategies (Count, Social Connectedness Index (SCI) an distance-aware Social Connectedness Index (DSCI)). *(B)* Distance dependence of connection probabilities in the OSN and RSN *(C)* Correlation of DSCI edge weights for different distance ranges

weights by the population size of the edge endpoints. Focusing on the results for SCI in Fig. 2A, we observe slightly lower but still high similarities compared to the plain edge weights: 0.918 for the combined layers, 0.883 for the family, 0.784 for the school, and 0.873 for the work layers. This suggests that the edge weights of the two networks are similar beyond population distribution patterns. Again, combining layers gives the highest similarity.

However, these correlations might still be partially driven by distance dependence, namely that closeby places are more likely to be connected. This relationship is often formalized as a gravity law (Lambiotte et al. [29]), in which the connection probability between areas has a power-law dependence on the Euclidean distance with a negative exponent – sometimes exactly −2 as in the original gravity law, but mostly ranging between −0.5 to −1.6 for various different social networks (Krings et al. [26], Scellato et al. [49], Bokányi et al. [5], Lengyel et al. [34], Xu et al. [60]). Inspecting Fig. 2B that shows the connection probability between two municipalities as a function of their Euclidean distance, we can observe a strong distance decay. The distance decay is slower for the OSN, which indicates more large-distance connections that are not as sensitive to distance as compared to the RSN. It should be noted that the overall higher probability of the OSN edges at all distances can be attributed to normalizing the probabilities by user counts instead of population size.

As a next step, we account for the distance dependence of the connection probability and propose DSCI: a distance-aware SCI metric that beyond population, is also normalized by the expected SCI at a certain distance. Thus, this metric reflects the similarity in tie formation disregarding the effects of both population size or spatial distance. Figure 2A shows that DSCI correlations between the OSN and RSN drop significantly compared to the plain SCI measure. The similarity of the OSN and the combined layers of the RSN

drops to 0.705. This is in the same range as for the family layer, which is 0.676, but we see smaller similarity scores for the school layer (0.383) and the work layer (0.548).

Because the OSN has more large-distance connections, we investigate how the correlation changes if we restrict the calculations to edges of different distances in Fig. 2C. We find that the correlation between DSCI edge weights decreases as the distance increases. Thus, the two networks compare best at local spatial scales, due to the OSN having more long-distance ties.

We also calulcate the Pearson correlations of the node-level metrics for the two additional networks: SCI weights and DSCI weights. The distribution, scatterplots and the exact values of the correlations for weighted degree, clustering coefficient, and pagerank can be found in the Additional file 1 Fig. 1. Correlation values are only slightly decreased for these normalized weights, ranging from 0.823 to 0.900, indicating a high level of structural similarity between nodes regardless of the edge weight normalization method.

2.3 Multivariate modeling

To show whether the OSN edge weights are significant even in the presence of the control variables population and geographical distance for the modeling of RSN edge weights, we ran multivariate OLS models. The results are summarized in Table 1 in the Additional file 1. For all three edge weighting strategies, we first include a baseline which models the RSN edge weights (log) with the population (log) and the geographic distance (log) of the two municipalities, along the lines of a gravity-like spatial interaction model (Illenberger et al. [21], Lee et al. [33]). By taking the logarithm of the dependent and independent variables, the obtained coefficients correspond to the exponents in a gravity-like model formulation. The sample used for the estimation are all municipality pairs with at least 10 edges in both the OSN and the RSN, self-loops (i.e. municipality pairs where the two municipalities are the same) excluded, corresponding to 78,947 observations.

In the baseline models (1,3, and 5), for raw edge counts, the population coefficients are positive and significant, and the distance coefficient is negative and significant, in line with the gravity law expectations of increasing social connectivity with population, and the decreasing connectivity with increasing geographical distance. In baseline models for SCI and DSCI, we see the coefficients of those variables which are included into the edge weight normalizations diminishing.

In three subsequent models (2,4, and 6), we add the logarithm of OSN edge weights to the baseline models. The coefficient for the OSN edge weights is always positive and significant, moreover, the predictive power of all three models, as measured by the R^2, increases – the increase is especially remarkable in the case of DSCI, from 0.026 in model (5) to 0.508 in model (6). This suggests that the edge weights of the OSN are good predictors of the RSN edge weights, and this predictive power goes beyond the gravity law, further confirming the similarity between the two networks.

2.4 Community structure

In this section, we explore the community structure of the municipality-level OSN and the RSN using two different community detection methods, and compare our results across algorithms, in addition to administrative boundaries and two distinct socio-cultural units. This allows us to reveal commonalities in the patterns of the community structure of the OSN ad the RSN. As far as we are aware, this is the first community analysis of the entire

social structure of the Netherlands. Intuition and public discourse would suggest a distinct community centered around the urbanized Randstad area in the west (with Amsterdam, The Hague, and Rotterdam), separate from the North and the South (Kooij and van de Laar [24]).

We use the Louvain modularity optimization algorithm (Blondel et al. [3]) combined with consensus clustering (Kwak et al. [28], Mandaglio et al. [37]) as well as a Stochastic Block Model (SBM) approach (Peixoto [46]) to obtain network communities. For further explanation see the Data and methods section. The pairwise similarities of the discovered communities are calculated using the adjusted mutual information (AMI) metric. An AMI value of 1 means identical partitions and 0 means that the partitions are only as similar as expected due to random chance.

Figure 3 presents our community finding results: it shows the maps of Louvain communities and the various cross-comparisons of the Louvain and SBM communities. In the tables of Fig. 3A-C, top left 4×4 corner displays the AMI similarity scores comparing the OSN and RSN communities. Figure 3A is based on the edge weights representing connection counts, Fig. 3B uses the SCI, Fig. 3C the DSCI metric. The similarity between the two networks' communities using the same algorithm (Louvain or SBM) is high for the counts (0.89 and 0.72) and the SCI metric (0.76 and 0.82). The results from the two community finding algorithms for the same network, however, display lower similarities (i.e. the AMI score for the OSN Louvain and SBM on Fig. 3A is 0.31). This is partly attributable to the different number of found communities: 4 to 7 with the Louvain method, and 8 to 14 with the SBM method. The country maps of Figs. 3A and 3B displaying the Louvain community detection results show spatially contingent community structures. This highlights a localized geographical preference in group formation. The maps of the SBM communities can be found in Additional file 1 Fig. 6.

The 'Province' row and column in the tables of Fig. 3A-C indicate the similarity of the resulting partitions from community detection to pre-established administrative boundaries, in our case, the subdivision of the Netherlands into 12 provinces. We highlight the borders of these provinces in Fig. 3A-C in white. We see that some communities largely follow province borders, with only small deviations, whereas other clusters span multiple provinces. Numerically, the first two edge weighting strategies (Count and SCI) combined with the Louvain algorithm give AMI scores for both networks between 0.61 and 0.73. Corresponding SBM communities, however, are less similar to provinces (0.30-0.32).

If we normalize edge weights for distance as well, i.e., look at DSCI in Fig. 3C, we get a lower similarity, 0.47, between the two Louvain community structures, but still 0.66 between the SBM community structures. DSCI communities have lower AMI scores for the Louvain–Province comparison, and higher AI scores for the SBM–province comparison in contrast with the Count and SCI communities. The number of communities in the SBM (14) is closer to the number of provinces (12), thus, both the lower values for Louvain, and the higher ones for SBM indicate that the administrative divisions are partly present in the communities found by the two algorithms.

The visual representation of the DSCI Louvain results on the map suggests that the communities are no longer spatially contiguous. We find one specific community in both networks that spans from the southwest to the northeast. This roughly follows the area of the Bible Belt of the Netherlands, a set of regions that form a distinct sociocultural unit with high shares of religious adherence and conservative voters (Exalto and Bertram-

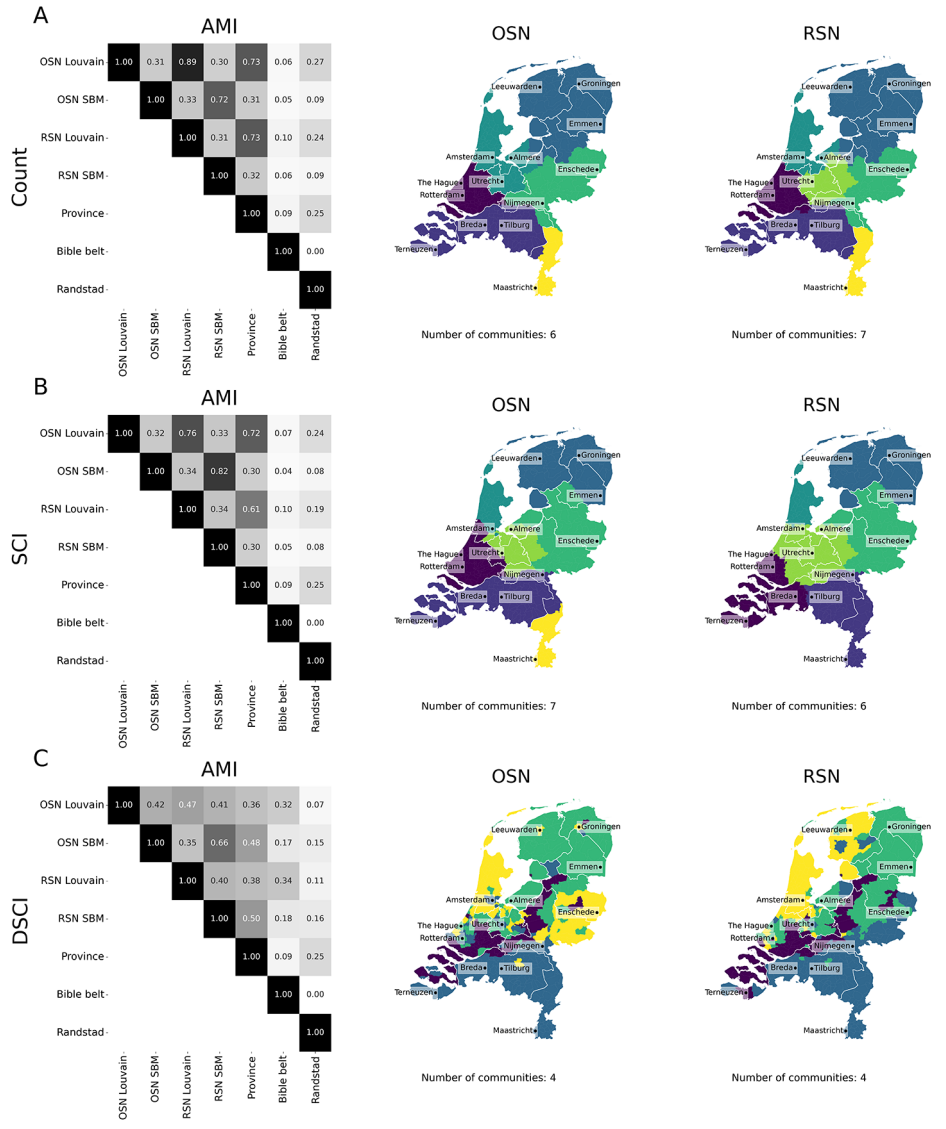

Figure 3 *Comparing community structures of the OSN and the RSN.* Community structure similarity of the online social network (OSN), the register-based social network (RSN), administrative borders (Provinces), and the Bible Belt and Randstad areas, measured in terms of Adjusted Mutual Information (AMI) for *(A)* the plain weighted network, *(B)* weights normalized by population density (SCI) and *(C)* population density and distance (DSCI) for Louvain and Stochastic Block Model (SBM) community detection algorithms. The maps in the figure represent the Louvain communities, for SBM communities see Additional file 1 Fig. 6

Troost [14]). To make the community finding results numerically comparable to the definitions of the Bible Belt and our expectations about the Randstad, we categorized municipalities based on whether they formally belong to these areas, and compared the categorization with the obtained partitions. The Bible Belt is defined as the set of municipalities in which the popular vote for the SGP party (Staatskunding Gereformeerd Partij – a conservative Christian party) exceeds the 4% threshold (Rellstab [47]). While there is no formal definition for the Randstad municipalities, they can be approximated with the urbanized municipalities of the four provinces Noord-Holland, Zuid-Holland, Utrecht, and Flevoland (Kooij and van de Laar [24], Zonneveld and Nadin [61]). We created addi-

tional figures in the Additional file 1 (Additional file 1 Fig. 8 and 9) representing overlaps of the clusters with the Bible Belt and the Randstad. Especially the Bible Belt municipalities are dense in a few communities in the DSCI results of both the SBM and Louvain algorithms. We also added AMI scores in Fig. 3 comparing the binary partitioning of Bible belt / non-Bible belt or Randstad / non-Randstad to the community finding results in last two columns of the AMI tables. We observe and increased AMI of the DSCI communities for both the Louvain and the SBM results, 0.17-0.34, as compared to the previous weighting stragegies 0.04-0.10, numerically confirming that the communities based on DSCI weighting highlight these two sociocultural areas. Again, a slightly lower score is attributable to the differing number of partitions in the communities and the binary partitioning corresponding to the Bible belt and the Randstad.

3 Discussion

It is well-known in the social network analysis literature that each source of social network data comes with its particular biases, as well as data completeness and data quality issues (Lazer et al. [31]). In the following, we expand on the usability and potential limitations of RSNs in the light of our findings.

We have analyzed the similarities between two Dutch large-scale networks from different data sources: Hyves, an online social network (OSN), and a register-based social network of Statistics Netherlands (RSN), each aggregated to the municipality level. We found that the two networks are very similar in terms of the topology of the nodes (degree, clustering coefficient, and pagerank) and connectivity. The RSN is most similar to the OSN when including all available edge types, thus, when combining multiple contexts of life in the construction of the register-based network. Moreover, the OSN weights are significant and improve the predictive power of a simple multivariate gravity model when modeling the RSN edge weights.

This suggests that behind the 'friend' edges of online social networks, there are multiple different mechanisms at play when users establish connections. While OSNs do a good job of capturing the social opportunity structure as established by the RSNs (and the other way around), OSNs are composed of several undisclosed edge sets with different characteristics. This result is in line with the literature that suggests that a large share of informal relationships are based on various forms of current or former family, school, and work connections of people (Van Eijk [58]). This also implies that even though a register-based network does not capture informal connections by definition, a superposition of various formal connections performs well when modeling an aggregated social network of a whole country.

We postulated that the very high similarity between the two networks can be by the simple fact that the same main factors of edge formation drive the connectivity in both networks. Therefore, we performed two normalization approaches: we calculated the SCI, a population-normalized version of the edge weights, and the DSCI, the deviation from the expected SCI at a given geographical distance. We find that the Pearson correlation of the edge weights decreases if we apply the normalization. However, it remains at a relatively high value of 0.7 even for DSCI. Thus, there remains a considerable similarity in social connectivity even after accounting for the population density and distance effects. This is further supported by the evidence of a multivariate statistical model, in which we find that OSN edge weights (for all weighting strategies) are significant and improve predictive power for modeling RSN edge weights even when controlling for population and distance.

The remaining similarity when using the population and distance-normalized weight DSCI most likely captures socio-cultural homophily and geographical or economic constraints, footprints of which are in the structure of both types of municipality-level networks. This is reflected in the multivariate statistical model where explanatory power of the gravity model is very low, and it increased considerably when adding the OSN DSCI weights to the model, thus indicating that similar remaining edge formation mechanisms underlie both networks. The fact that the correlation of DSCI edge weights decreases with increasing distance highlights that mostly the local structures are similar in the OSN and the RSN. Long-distance connections are both more numerous and more random in the OSN used in this paper, which we can partly trace back to a limitation in the original underlying data, namely, the sampling strategy for work relationships – for companies larger than 100, the 100 geographically closest colleagues are chosen as work connections for a person. This sampling is designed to overcome the problem that the Dutch administrative registers do not contain information about the exact location of the employee within an organization. The sampling is reflected in both the faster distance decay of the RSN edge probability, and the lower correspondence between the two networks for larger distances. While the choice of including the 100 closest colleagues into the work-related social opportunity structure of individuals seems to be an arbitrary limit, a random sampling approach would include colleagues from faraway subunits with even lower chance of an actual realized relationship than the distance-based sampling. The distance-decay behaviour of family and school RSN edges (Additional file 1 Fig. 7), however, suggests that if better quality work data would be available, and we would not have to rely on a sampling strategy for large organizations, it would improve the RSN representation of large-distance connections.

Our first analysis of the community structure of the whole of the Netherlands also revealed striking similarities between the OSN and RSN. Again, we find the least correspondence between the OSN and RSN communities when using normalization for both distance and population. However, in this latter case, a remarkable community that is not geographically contiguous appears in both the OSN and RSN. It includes most of the so-called Dutch Bible Belt, which is a set of regions that form a distinct sociocultural unit with high shares of religious adherence and conservative voters (Exalto and Bertram-Troost [14], Rellstab [47]). Thus, distance- and population-aware communities uncover socioculturally similar regions in both networks in a similar way to Expert et al. [15]. If we compare communities with the administrative boundaries of Dutch provinces, we find that distance-unaware communities somewhat correspond to administrative boundaries. However, there are notable differences, such as the northern part of the province of Flevoland attaching to Overijssel. This might reflect that economic and infrastructural constraints matter more than administrative division in this case, since the Northern parts of Flevoland are infrastructurally well-connected to the neighboring province. When using DSCI, there is very little agreement between network communities with province boundaries. Hence, socioeconomic policymaking on topics such as labor markets, infrastructure investments, or formal care systems might be better based on community clusters rather than provincial boundaries.

The differences in the structures of the OSN and RSN highlighted throughout this paper might originate from the different underlying link generation strategies in the two networks. In the RSN, family links represent persistent ties, whereas work and school re-

lationships only reflect the situation of the current year. OSNs capture the fact that some relationships are always retained from former schools or workplaces, even if sometimes in a different context such as a close friendship. Aggregating RSN links over time (e.g. similarly to Cremers et al. [10]) and comparing them with the OSN structure could provide further insight into this matter. However, OSN links might reflect connections of very different strengths, ranging from close family to distant past acquaintances. This can partly explain the differences in the structure of large-distance connections between the RSN and the OSN. RSNs miss out on important informal relationships such as church or leisure groups. The OSN node sample is also heavily biased towards younger users (Corten [9]), which may have important consequences on both the number of links – young adults tend to have the highest number of connections (Bokányi et al. [4]) –, and the distance distribution of those edges if younger people are more mobile. We observe the structural similarity of the two municipality-level networks despite these differences.

4 Conclusion

In this work, we provided an comparison of the population-scale network structure of an online social network (OSN) and a register-based social network (RSN) in the Netherlands. We observed similarities between the structure of the two networks despite the OSN containing self-reported friendship ties, and the RSN being based on legal definitions of kinship and formal affiliations such as work and school. We showed that the two networks are similar when comparing structural metrics of modes, their connectivity; that combining all available RSN layers (family, school, work) results in the highest similarity; that similarity remains relatively high even after accounting for population and spatial distance patterns, especially for local edges; and that OSN edge weights are significant and improve predictive power when modeling RSN edges.

By analyzing communities of the two networks with different edge weighting strategies with multiple algorithms, we showed that the networks have similar community structures using all three edge weighting strategies; and that detected communities do not closely follow pre-established administrative borders, especially when accounting for population and distance patterns. However, the latter method uncovers a socioculturally tightly knit community that corresponds to the Dutch Bible Belt.

In summary, we expect researchers to draw similar conclusions based on register-based social networks and online social networks. Both data sources are useful for modeling the social network structure of a whole population, and the more edge types a register-based network contains, the better the comparability.

5 Data and methods

In this section, we first present Hyves and the register-based social network of Statistics Netherlands, the OSN and RSN datasets, and their aggregation into Dutch municipalities, followed by a detailed description of the RSN layers. Then, we introduce our notation and describe the methods for normalizing edge weights. Lastly, we outline the process of identifying communities.

5.1 Social network datasets

OSN. The Hyves online social network was an online social media platform in the Netherlands (Corten [9]) before the advent of Facebook. The dataset represents the late 2009 state

of the network which during its peak period contained 10M people, covering up to 60% of the population of the Netherlands. The network represents supposed friendship connections between its registered users. There are 6.2M users with a self-reported place of residence at a municipality-level resolution, with 320M edges between them. We excluded users flagged as celebrities from our analysis.

The self-reported municipality names in the data were provided by users and therefore were prone to different errors. Norbutas and Corten [42] cleaned and aggregated place names at the municipality level even if users gave different administrative units, such as neighborhoods, as their place of residence. The municipalities were matched to the official list of Statistics Netherlands as of 2009. We use this cleaned and matched municipality dataset, and refer the reader to Norbutas and Corten [42] for more details on data processing.

RSN. The register-based social network (RSN) is compiled by Statistics Netherlands (CBS) from official records (van der Laan and de Jonge [57], van der Laan [55]). In this network, the nodes are all 16.6M residents of the Netherlands in 2010. The roughly 570M edges are organized in three layers representing different contexts of life: family, school and work relations. Family relations are derived from the official parent-child and partner registers, school relations are projections of the bipartite person-school network sourced from educational records, and work relations are derived based on the bipartite employer-employee matching. Each person's place of residence (municipality) is known. In the following, we give more details on the different layers.

Family layer. Family connections are derived from official parent-child and partner relations. The partner relations are united from marriage registers, tax declarations, and household registers. From the above two source datasets, other family ties such as grandparents, grandchildren, siblings, aunts, uncles, cousins, nieces, and nephews are inferred. Step and in-law relationships are also included.

School layer. School connections are aggregated from various official educational agencies spanning five levels of education: elementary, secondary, secondary special, vocational, and higher. People have a school connection if they go to the same school, in the same yeargroup, at the same school location (larger organizations might have multiple sites in different municipalities), and type of education (e.g. vocational or secondary). University and other higher education students are further distinguished by study programmes. Teachers are not linked to students through schools.

Work Work connections contain links between people working for the same employer of their major source of income. If a company has less than 100 employees, all of them are connected to each other. Otherwise, a person is connected only to the 100 co-workers closest to their residence.

We introduce the intuition and notion of the aggregated network of municipalities created from the person network. As direct person matching between the two networks is infeasible, we aggregate the networks such that nodes are the 431 municipalities of the Netherlands in 2010, and weighted edges between municipality pairs count the number of ties between people in the municipalities that the link connects. There were a few municipality merges in The Netherlands from 2009 to 2010 that we applied to the network dataset as well. We obtain this aggregated network for the OSN and every relevant layer in the RSN (family, school, and work), as well as for a combination of these layers. In the combined layers, an edge exists between two people when at least an edge exists in any of

the three layers. If multiple edges run between people in the base layers, we count it as a single edge in the aggregated layer.

5.2 Preliminaries

We introduce the notion of the aggregated multilayer graph. The notation is based on Bokányi et al. [4]. We represent a person-level network as $G_p = (V_p, E_p, L)$, where V_p is the set of nodes respresenting people. In our case, the residents of the Netherlands in 2010 consisted of $|V_p| = n_p = 16.6$M people. The set of undirected edges running between these nodes can be described as

$$E_p \subseteq \{(\{u, v\}, \ell) : u, v \in V_p,\ u \neq v,\ \ell \in L\}, \tag{1}$$

such that L is the set of possible layers. In this setting, we can represent the network G_p using personal level binary adjacency tensor $(A_p)_{u,v,\ell}$. An entry $a_{u,v,\ell}$ of this matrix is 1 if and only if an edge runs between persons $u, v \in V_p$ in layer $l \in L$, and 0 otherwise.

We define $G = (V, E, L)$ as a multilayer graph. In this case, V is the set of municipalities, $n = |V| = 431$ is the number of nodes. The set of undirected edges is

$$E \subseteq \{(\{u, v\}, \ell, w) : u, v \in V,\ u \neq v,\ w \in \mathbb{R},\ \ell \in L\}, \tag{2}$$

such that L is the set of the possible layers and w is the strength of the connection between the two municipalities, for which we propose three different weighting schemes in the Weighting strategies section below. We can represent the edges E using an adjacency matrix $A_{u,v,\ell}$ that counts half-edges that run between $u, v \in V$ in the layer $\ell \in L$.

We can relate the two representations as follows. We can represent the place of residence using a binary affiliation matrix B of shape $n_p \times n$. An entry in this matrix is 1 if $u \in V_p$ is affiliated to $v \in V$. With the help of this representation, we can calculate

$$A_{u,v,\ell} = B^T (A_p)_{u,v,\ell} B. \tag{3}$$

Here, $(.)^T$ represents matrix transposition.

Person-level edges can also originate and end in the same municipality. This is represented by weighted self-edges in the aggregated graph. The above equation would count person-level edges twice within the same municipality. However, we dropped all of the self-edges when running our experiments.

In this weighted municipality-level network, the edge weight w_ij thus represents connection counts. When connections from all layers are added,

$$w_{ij} = \sum_{\ell} A_{u,v,\ell}.$$

5.3 Weighting strategies

We compare the strengths of the connections between municipality pairs in the RSN and the OSN by comparing edge weights corresponding to the same municipality pair. We incorporate population and distance into the weighting scheme as follows.

Population corrected weighting. We use a metric that not only counts connections between areas but also takes into account that larger population areas typically have more

connections inspired by Bailey et al. [2]. Within a layer, the metric between $i, j \in V$ is formulated as

$$SCI_{ij} = \frac{\text{Connections}_{ij}}{\text{Possible connections}_{ij}}. \tag{4}$$

If $i \neq j$, then the number of possible connections is $\text{Population}_i \times \text{Population}_j$. Otherwise, it is equal to $\text{Population}_i \times (\text{Population}_j - 1)$. In the case of the RSN, the population is the number of inhabitants. In the case of the OSN, population is the number of users that have self-reported the municipality as location in their profiles.

Population and distance corrected weighting. It is widely known that distance is an important factor when forming connections (Lambiotte et al. [29]). A power-law distribution often models this dependency which is often called the gravity law. In our context, we use a model-free metric inspired by Expert et al. [15]. To measure the distances of municipalities, we calculate the distances of the centroids of the municipalities using the Euclidean distance metric.

The proposed distance-aware social connectivity index (DSCI) is given by

$$DSCI_{ij,D} = \frac{SCI_{ij}}{\mathbb{E}\left[SCI|D\right]}, \tag{5}$$

where D denotes a certain spatial distance and $\mathbb{E}[\cdot]$ denotes the expected value of a variable. We approximate this value by creating 200 bins that contain an equal number (464) of municipality pairs between 0 and 360 km.

5.4 Node-level features

We use three metrics from the `networkx` Python package to characterize nodes in the three weighted networks.

1. We calculate the weighted degree of nodes:

$$D_i^w = \sum_j w_i j.$$

2. We use the weighted version of the clustering coefficient which in the package corresponds to the geometric average of the subgraph edge weights (Onnela et al. [44]).
3. Third, we calculate the PageRank of each node to reflect its centrality with damping parameter $\alpha = 0.15$.

5.5 Community detection

Community detection (Newman [41], Girvan and Newman [18]) is a way to identify groups of nodes in a network that form tightly knit subunits that are more loosely connected with other subunits. We use this to investigate the meso-level structure in our networks. We perform community detection based on the Louvain method, which accounts for edge weights. This allows us to investigate the three community structures resulting from our three edge weighting strategies: the number of connections between municipalities, the SCI weights, and the DSCI weights. We set the resolution parameter to 1. We use the `Python` package `networkx` (Hagberg et al. [19]).

It is well-known that community detection algorithms involve a degree of randomization. This can be accounted for by using consensus clustering (Lancichinetti and Fortunato [30]). In our first set of experiments, we use 1) 1000 iterations of the Louvain algorithm to 2) create a new network based on the number of times the nodes belonged to the same community. Then, we go to step 1) and repeat until convergence. It took 3 iterations until all node pairs distinctively belonged to the same community. In the second set of experiments, we use the Stochastic Block Model (SBM) of Peixoto [46] implemented in the `graph-tool` Python package to find non-overlapping communities. The results in both cases can also be regarded as a partitioning of the node set V into R non-empty partitions $U = \{U_1, U_2, \ldots, U_R\}$, where $U_i \cap U_j = \{\}$ for any $i \neq j,\ i,j \in \{1,2,\ldots,R\}$, so the partitions are pairwise disjoint, and $\cup_{i=1}^{R} U_i = V$.

We use the Adjusted Mutual Information (AMI) metric of Vinh et al. [59] to compare the partitions we get from the concensus clustering and the SBM on the different networks and edge weighting strategies, and also to compare the network partitions to the province borders of the Netherlands, and the Bible belt and tha Randstad areas which in essence are also partitioning into 12, 2, and 2 partitions, respectively. If we have two different partitionings, $U = \{U_1, U_2, \ldots, U_R\}$ of R partitions, and $T = \{T_1, T_2, \ldots, T_C\}$ of C partitions, then the Adjusted Mutual Information is:

$$AMI(U,T) = \frac{MI(U,T) - \mathbb{E}\{MI(U,T)\}}{\text{avg}\{H(U),H(T)\} - \mathbb{E}\{MI(U,T)\}}. \tag{6}$$

In the above equation, MI stands for Mutual Information, which if calculated as

$$MI(U,T) = \sum_{i=1}^{R}\sum_{j=1}^{C} P_{UT}(i,j)\log\frac{P_{UT}(i,j)}{P_U(i)P_T(j)}, \tag{7}$$

where $P_{UT}(i,j) = \frac{|U_i \cap T_j|}{|V|}$ is the probability that a node belongs to partition i in U, and partition j in T, and $P_U(i) = \frac{|U_i|}{|V|}$ is the probability that a node belongs to partition i in U, and $P_T(j) = \frac{|T_j|}{|V|}$ is the probability that a node belongs to partition j in T. $\mathbb{E}$ denotes the expected value of the Mutual Information, for details on its calculations, we refer the reader to Vinh et al. [59]. The expected MI terms normalize this score to reflect that two random partitionings can also have similarity by chance.

H stands for the entropy associated with a partitioning U:

$$H(U) = \sum_{i}^{R} -P_U(i)\log P_U(i). \tag{8}$$

We use the implementation of the `scikit-learn` (Pedregosa et al. [45]) Python package (see `sklearn.metrics.adjusted_mutual_info_score`) for the calculations.

List of abbreviations

CBS, Statistics Netherlands; DSCI, distance-aware Social Connectedness Index; OSN, online social network; RSN, register-based social network; SBM, Stochastic Block Model; SCI, Social Connectedness Index.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-025-00522-4.

Additional file 1. (PDF 4.9 MB)

Acknowledgements
We would like thank the POPNET team (www.popnet.io) as well as the anonymous reviewers for helpful suggestions and discussions.

Author contributions
E.B., F.T., R.C., and E.H. conceptualized the study, M.M, E.B., and R.C. prepared the data, M.M., E.B., Y.K., and F.T. analyzed the data, M.M. and E.B. created the figures, M.M, E.B.,Y.K.,F.T, and E.H. wrote the manuscript.

Funding
The POPNET project has been funded by Platform Digitale Infrastructuur Social Sciences and Humanities (www.pdi-ssh.nl).

Data Availability
The aggregated municipality-municipality networks are available via https://doi.org/10.5281/zenodo.13859916. Under certain conditions, the underlying microdata (not directly analyzed or used in this work) are accessible for statistical and scientific research. For further information, contact the corresponding author or microdata@cbs.nl.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands. [2]University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. [3]Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands.

Received: 30 September 2024 Accepted: 10 January 2025 Published online: 28 January 2025

References

1. Asikainen A, et al (2020) Cumulative effects of triadic closure and homophily in social networks. Sci Adv 6(19):eaax7310. https://doi.org/10.1126/sciadv.aax7310
2. Bailey M, et al (2018) Social connectedness: measurement, determinants, and effects. J Econ Perspect 32(3):259–280. https://doi.org/10.1257/jep.32.3.259
3. Blondel VD, et al (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
4. Bokányi E, Heemskerk EM, Takes FW (2023) The anatomy of a population-scale social network. Sci Rep 13(1):9209. https://doi.org/10.1038/s41598-023-36324-9
5. Bokányi E, Novák M, et al (2022) Urban hierarchy and spatial diffusion over the innovation life cycle. R Soc Open Sci 9(5):211038. https://doi.org/10.1098/rsos.211038
6. Cardon PW, et al (2009) Online and offline social ties of social network website users: an exploratory study in eleven societies. J Comput Inf Syst 50(1):54–64. https://doi.org/10.1080/08874417.2009.11645362
7. Chetty R, et al (2022) Social capital I: measurement and associations with economic mobility. Nature 608(7921):108–121. https://doi.org/10.1038/s41586-022-04996-4
8. Chetty R, et al (2022) Social capital II: determinants of economic connectedness. Nature 608(7921):1–13. https://doi.org/10.1038/s41586-022-04997-3
9. Corten R (2012) Composition and structure of a large online social network in the Netherlands. PLoS ONE 7(4):e34760. https://doi.org/10.1371/journal.pone.0034760.
10. Cremers J, et al (2024) Unveiling the social fabric: a temporal, nation-scale social network and its characteristics. https://doi.org/10.48550/arXiv.2409.11099. (Visited on 09/25/2024). Pre-published
11. Dunbar RIM, Arnaboldi V, Conti M, Passarella A (2015) The structure of online social networks mirrors those in the offline world. Social Networks 43:39–47. https://doi.org/10.1016/j.socnet.2015.04.005
12. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328(5981):1029–1031. https://doi.org/10.1126/science.1186605
13. Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. Proc Natl Acad Sci USA 106(36):15274–15278. https://doi.org/10.1073/pnas.0900282106
14. Exalto J, Bertram-Troost G (2019) Strong religion in a secular society: the case of orthodox reformed schools in the Netherlands. Educ Sci 9(1):28. https://doi.org/10.3390/educsci9010028
15. Expert P, et al (2011) Uncovering space-independent communities in spatial networks. Proc Natl Acad Sci USA 108(19):7663–7668. https://doi.org/10.1073/pnas.1018962108

16. Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174. https://doi.org/10.1016/j.physrep.2009.11.002
17. Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659:1–44. https://doi.org/10.1016/j.physrep.2016.09.002
18. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99(12):7821–7826. https://doi.org/10.1073/pnas.122653799
19. Hagberg A, Swart PJ, Schult DA (2008) Exploring network structure, dynamics, and function using NetworkX. English. Tech. rep. LA-UR-08-05495; LA-UR-08-5495, Los Alamos National Laboratory (LANL), Los Alamos, NM (United States)
20. Hidalgo CA, Rodriguez-Sickert C (2008) The dynamics of a mobile phone network. Phys A, Stat Mech Appl 387(12):3017–3024. https://doi.org/10.1016/j.physa.2008.01.073
21. Illenberger J, Nagel K, Flötteröd G (2013) The role of spatial interaction in social networks. Netw Spat Econ 13(3):255–282. https://doi.org/10.1007/s11067-012-9180-4
22. Kassa YM, Cuevas R, Cuevas A (2018) A large-scale analysis of Facebook's user-base and user engagement growth. IEEE Access 6:78881–78891. https://doi.org/10.1109/ACCESS.2018.2885458
23. Kazmina Y, et al (2024) Socio-economic segregation in a population-scale social network. Soc Netw 78:279–291. https://doi.org/10.1016/j.socnet.2024.02.005
24. Kooij P, van de Laar P (2003) The randstad conurbation: a floating Metropolis in the Dutch delta. In: The European Metropolis 1920-2000, European Science Foundation
25. Kovács ÁJ, et al (2023) Income-related spatial concentration of individual social capital in cities. Environ Plan B: Urban Anal City Sci 50(4):1072–1086. https://doi.org/10.1177/23998083221120663
26. Krings G, et al (2009) Urban gravity: a model for inter-city telecommunication flows. J Stat Mech Theory Exp 2009(7):L07003. https://doi.org/10.1088/1742-5468/2009/07/L0700310.1088/1742-5468/2009/07/L07003
27. Kumar R, Novak J, Tomkins A (2006) Structure and evolution of online social networks. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 611–617
28. Kwak H, et al (2011) Consistent community identification in complex networks. J Korean Phys Soc 59(5):3128–3132. https://doi.org/10.3938/jkps.59.3128
29. Lambiotte R, et al (2008) Geographical dispersal of mobile communication networks. Phys A, Stat Mech Appl 387(21):5317–5325. https://doi.org/10.1016/j.physa.2008.05.014
30. Lancichinetti A, Fortunato S (2012) Consensus clustering in complex networks. Sci Rep 2(1):336
31. Lazer D, Hargittai E, et al (2021) Meaningful measures of human society in the twenty-first century. Nature 595(7866):189–196. https://doi.org/10.1038/s41586-021-03660-7
32. Lazer D, Pentland A, et al (2009) Computational social science. Science 323(5915):721–723. https://doi.org/10.1126/science.1167742
33. Lee C, Scherngell T, Barber MJ (2011) Investigating an online social network using spatial interaction models. Soc Netw 33(2):129–133. https://doi.org/10.1016/j.socnet.2010.11.002
34. Lengyel B, et al (2015) Geographies of an online social network. PLoS ONE 10(9):e0137248
35. Liben-Nowell D, et al (2005) Geographic routing in social networks. Proc Natl Acad Sci USA 102(33):11623–11628. https://doi.org/10.1073/pnas.0503018102
36. Magnani M, et al (2022) Generation of temporal multilayer networks from population-scale register data. In: The 6th European conference on social networks
37. Mandaglio D, Amelio A, Tagarelli A (2018) Consensus community detection in multilayer networks using parameter-free graph pruning. In: Advances in knowledge discovery and data mining: 22nd Pacific-Asia conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, proceedings, part III. Springer, Melbourne, pp 193–205. https://doi.org/10.1007/978-3-319-93040-4_16
38. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Annu Rev Sociol 27(1):415–444. https://doi.org/10.1146/annurev.soc.27.1.415
39. Mislove A, et al (2011) Understanding the demographics of Twitter users. In: Int'l AAAI conference on weblogs and social media (ICWSM), pp 554–557
40. Myers SA, et al (2014) Information network or social network? The structure of the Twitter follow graph. In: Proceedings of the 23rd international conference on world wide web. WWW '14 companion. ACM, New York, pp 493–498. https://doi.org/10.1145/2567948.2576939
41. Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103(23):8577–8582. https://doi.org/10.1073/pnas.0601602103
42. Norbutas L, Corten R (2018) Network structure and economic prosperity in municipalities: a large-scale test of social capital theory using social media data. Soc Netw 52:120–134. https://doi.org/10.1016/j.socnet.2017.06.002
43. Onnela J-P, Saramäki J, Hyvönen J, et al (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci USA 104(18):7332–7336. https://doi.org/10.1073/pnas.0610245104
44. Onnela J-P, Saramäki J, Kertész J, et al (2005) Intensity and coherence of motifs in weighted complex networks. Phys Rev E, Stat Nonlinear Soft Matter Phys 71(6):065103
45. Pedregosa F, et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12(85):2825–2830
46. Peixoto TP (2018) Nonparametric weighted stochastic block models. Phys Rev E 97(1):012306. https://doi.org/10.1103/PhysRevE.97.012306
47. Rellstab S (2023) Gender norms and the child penalty: evidence from the Dutch bible belt. Appl Econ 56(45):5428–5441. https://doi.org/10.1080/00036846.2023.2257030
48. Savcisens G, et al (2024) Using sequences of life-events to predict human lives. Nat Comput Sci 4(1):43–56. https://doi.org/10.1038/s43588-023-00573-5
49. Scellato S, et al (2011) Socio-spatial properties of online location-based social networks. Proc Int AAAI Conf Web Soc Media 5(1):329–336. https://doi.org/10.1609/icwsm.v5i1.14094
50. Sloan L, et al (2013) Knowing the tweeters: deriving sociologically relevant demographics from Twitter. Sociol Res Online 18(3):74–84. https://doi.org/10.5153/sro.3001
51. Soler N, Heemskerk E, Kazmina Y (2024) Contacts in contexts: a population-scale social network approach to the study of close intergroup social ties. https://doi.org/10.31235/osf.io/axumt. (Visited on 05/28/2024). Preprint

52. Subrahmanyam K, et al (2008) Online and offline social networks: use of social networking sites by emerging adults. J Appl Dev Psychol 29(6):420–433. https://doi.org/10.1016/j.appdev.2008.07.003
53. Tóth G, et al (2021) Inequality is rising where social network segregation interacts with urban topology. Nat Commun 12(1):1143. https://doi.org/10.1038/s41467-021-21465-0
54. Ugander J, et al (2011) The anatomy of the Facebook social graph. https://doi.org/10.48550/ARXIV.1111.4503
55. van der Laan J (2022) A person network of the Netherlands. https://www.cbs.nl/-/media/_pdf/2022/20/person_network_netherlands.pdf
56. Van der Laan J, et al (2023) A whole population network and its application for the social sciences. Eur Sociol Rev 39(1):145–160. https://doi.org/10.1093/esr/jcac026
57. van der Laan J, de Jonge E (2017) Producing official statistics from network data. In: The 6th international conference on complex networks & their applications - book of abstracts, pp 288–290. https://doi.org/10.5281/zenodo.8019396
58. Van Eijk G (2010) Unequal networks: spatial segregation, relationships and inequality in the city. In: Sustainable urban areas, vol 32. Delft Univ Press, Delft. 352 pp.
59. Vinh NX, Epps J, Bailey J (2009) Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th annual international conference on machine learning. ICML '09. ACM, New York, pp 1073–1080. https://doi.org/10.1145/1553374.1553511
60. Xu Y, Santi P, Ratti C (2022) Beyond distance decay: discover homophily in spatially embedded social networks. Ann Assoc Am Geogr 112(2):505–521. https://doi.org/10.1080/24694452.2021.1935208
61. Zonneveld W, Nadin V (eds) (2020) The randstad: a polycentric Metropolis Routledge, London. https://doi.org/10.4324/9780203383346. 382 pp.

Publisher's Note