



OPEN

Quantifying social segregation in large-scale networks

Bjørn-Atle Reme^{1✉}, Andreas Kotsadam², Johannes Bjelland³, Pål Roe Sundsøy³ & Jo Thori Lind⁴

We present a measure of social segregation which combines mobile phone data and income register data in Oslo, Norway. In addition to measuring the extent of social segregation, our study shows that social segregation is strong, robust, and that social networks are particularly clustered among the richest. Using location data on the areas where people work, we also examine whether exposure to other social strata weakens measured segregation. Lastly, we extend our analysis to a large South Asian city and show that our main results hold across two widely different societies.

Similar people are more likely to form social ties. This phenomenon, referred to as social homophily, has been documented in several academic disciplines^{1–4}. Social homophily is typically based on socially salient characteristics, such as social class, gender, ethnicity, religion, or beliefs and leads to social segregation. Given the importance of social contact for building trust, empathy, and cooperation in a population^{5–10}, measuring social segregation is crucial. However, it is difficult as it requires data on the patterns of interpersonal contact in a population. With a few notable exceptions^{11–13}, most studies have therefore either proxied social segregation with spatial segregation, or studied social networks in smaller groups^{14–18}. While spatial segregation is important, it is insufficient as a proxy for larger scale social segregation. Most importantly, spatial proximity does not imply contact and studies show that closeness without contact may actually increase social divides^{14,19–22}.

Technological progress, and particularly the introduction of the smartphone, has allowed for utilizing high granularity mobile phone data. This has also led to a renewed interest in social network patterns^{19,23–30}. However, thus far, there have not been studies which combine detailed mobile phone records with income register data to study social segregation, while controlling for spatial proximity. This is the purpose of our study. We make several contributions to the literature on socioeconomic segregation. First, by combining mobile phone data and income register data, we quantify social segregation across income in Oslo, Norway. In particular, we estimate the association between income differences and communication intensity, while controlling for spatial proximity. Second, we examine how daytime exposure to other social strata affects the degree of social segregation. Third, we estimate how social clustering varies across income. Last, by extending the analysis to a large south-east Asian city, we explore the similarities of these patterns across widely different societies.

In contrast to self reported survey data, mobile data reflects actual behavior. In addition, by also including detailed data on spatial proximity in our analyses, this study offers an unprecedented view of actual social segregation. As economic opportunities, such as job market outcomes, are strongly influenced by social networks this segregation may have important consequences for vulnerable groups¹.

Data and methods

Measuring communication. To measure social contact in Oslo we utilize detailed call data records over a 3 month period in 2013 from the market's largest mobile network operator. This operator has close to 250,000 subscribers in the city area, covering approximately 50% of the population. We define a communication event as any communication in terms of phone call or text message. During the period at hand the subscribers initiated 36 million communication events. While internet-based services for messaging and calling, such as Skype, WhatsApp, Messenger and FaceTime, currently are the primary mode of communication with friends for many, this study is from 2013, a time when the penetration of such services was very limited. For example, a survey for this time period found that among the general population, more than 50% had never used such services, more than 20% used them less than once a month, and approximately 15% used only them 1–3 times per month³¹.

¹Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway. ²Ragnar Frisch Centre for Economic Research, Oslo, Norway. ³Telenor Research, Oslo, Norway. ⁴Department of Economics, University of Oslo, Oslo, Norway. ✉email: Bjorn-Atle.Reme@fhi.no

Measuring socioeconomic status. In order to measure socioeconomic status, we use data on labor earnings from the Norwegian register data, where we have reported earnings for all Norwegian citizens, as well as demographic details.

Linking communication and socioeconomic status. The study is in accordance with relevant guidelines and regulations. Therefore, due to privacy protection regulations, the call data records and income data were not linkable on the level of the individual. To link these data sources we aggregate the analysis on the level of a mobile base station, of which there are 689 in the sample (see Fig. A.3 in the Supplementary Material for maps showing mobile towers and communication in Oslo). Every cell phone subscriber is associated a *home tower*—the base station most used by the subscriber between 7:00 pm and 7:00 am in this 3 month period. The socioeconomic status of a home tower's catchment area is defined as the mean income of residents with positive incomes living in the area (see Table A.1; Fig. A.1 in Supplementary Material for descriptive statistics of the sample).

Measuring social segregation. The core of our analyses is based on the income and amount of communication between the 474,721 directed pairs of home towers (see Supplementary material Fig. A.2 for the distribution of between tower communication intensities). As our data are on communication between pairs of towers, each with a heterogeneous population of subscribers, and the economic differential between each pair of individuals is continuous, we rely on techniques studying the relationship between aggregate link strengths and aggregate differences in population composition as in¹³.

We assign an average income to each cell tower. We use register data on pre-tax wage for all individuals in Oslo in 2010. We know the basic unit ('grunnkrets') of residence of each individual. We can then average over all individuals residing in basic unit g to find the average income in that unit, y_g . To map incomes to cell towers, we construct cell tower t 's coverage by dividing the city into Voronoi polygons³². Let A_{tg} denote the area of overlap between basic unit g and tower Voronoi polygon t . If they don't overlap, $A_{tg} = 0$. Then the estimated income of residents of tower t 's catchment area is estimated as

$$\bar{y}_t = \frac{\sum_g A_{tg} y_g}{\sum_g A_{tg}}.$$

To summarise our approach, consider the following illustrating example: if an agent identified with tower t makes a call from a location associated with tower u , then the link is registered as emerging from tower t . Moreover, if the receiver lives at location v , but receives the call at location w , then the link is registered as between location t and v .

In the analysis of group specific incomes, we use the register data to compute age group and gender specific income averages and we use a similar formula for estimating the average income for this demographic group in tower t 's catchment area. One potential challenge with using group averages instead of individual level data is that it can lead to an aggregation bias: the group-level associations misrepresent the individual-level associations³³. To check the sensitivity of our results with respect to this type of bias, we have undertaken analyses where we impute group averages on income for 12 demographic groups—gender and six age groups. The results indicate that our primary income measure is best at explaining communication patterns (see Supplementary material Table A.5).

Asian data. We used 1 month of raw call data records from the country's largest carrier to construct a country-wide call graph. The number of total subscribers were 113 million, 2.7 billion communication links and 10,000 mobile towers. This dataset was further subset to only contain the links for the largest city—covering 18 million subscribers, 111 million social ties and 2974 towers.

For income we used survey data, since no reliable income register data exists. The income categories for a random selection of 76,005 subscribers were obtained through two sequential large-scale market research household surveys. Information about income was directly asked from the respondents, who were requested to place themselves within pre-defined income bins. The survey also contains geo-coordinates of the location of residence of each respondent. Respondents within the household were selected via the Kish grid method among those who were eligible³⁴. The correlation between the average income per region based on the survey results and their values published in official statistics were 0.925.

In order to calculate the income distribution at the cell tower level, the dataset was first restricted to contain participants from the largest city. Then income was aggregated on the tower level by assigning every respondent to her closest tower (by bird's flight).

Measuring correlations with income differences. To estimate the correlation between income differences and communication intensity, we estimate the regression model

$$\text{Events}_{ij} = \alpha + \beta |\ln \bar{y}_i - \ln \bar{y}_j| + \theta' z_{ij} + v_{ij}$$

Here Events_{ij} is the number of communication events between towers i and j , \bar{y}_i and \bar{y}_j the average income of residents at each tower, z_{ij} a vector of controls and v_{ij} the residual. The vector of covariates z_{ij} includes a fourth order polynomial specification of the geographical distance between the cell towers, the logarithm of the income level of the sending and receiving towers, the total tower traffic level of the sending and receiving tower and the expected tower traffic level. The effect of increased income differences on communication intensity is the parameter β . We also consider fixed effects specifications where the constant term α is replaced by sending and receiving specific intercepts $\alpha_i^S + \alpha_j^R$.

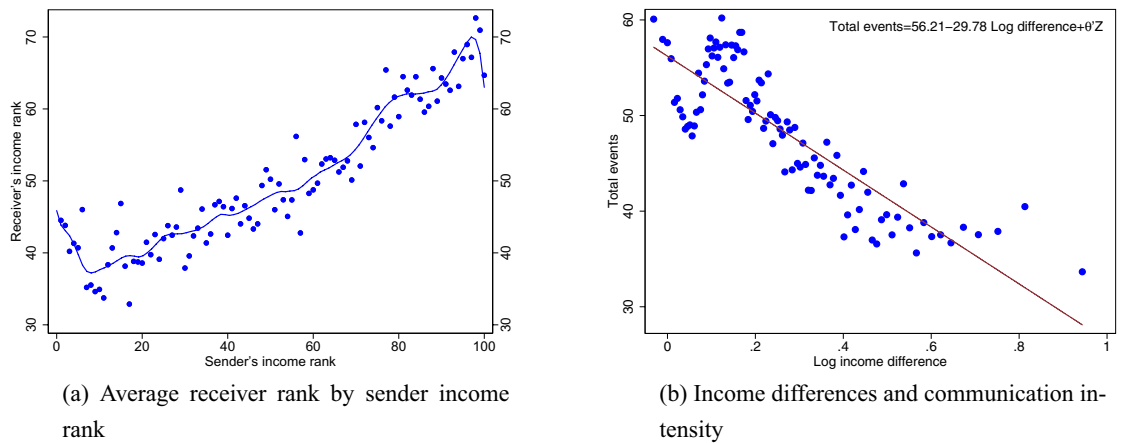


Figure 1. Economic segregation in telecommunication. The figure illustrates economic segregation in telecommunication in Oslo. Panel (a) shows the average income rank of receivers by the income percentile of the sender, as well as local linear smoothing of the relationship. Within tower communication is disregarded to exclude within household communication. Panel (b) shows a binned scatter plot with the number of communication events between cell phone towers against the absolute value of difference in log income between the two. Z includes controls for geographical distance between two cell towers (up to fourth polynomial), the income level of sending and receiving tower, total tower traffic level and expected tower traffic level.

Measuring the correlation with daytime exposure. An individual residing at a rich nighttime tower and frequenting a poor daytime tower, or vice versa, is said to be exposed to other socioeconomic groups. To measure daytime exposure to other social strata, we first identify each individual's most used cell tower between 12 am and 2 pm in the weekdays, the *daytime tower*. We then regress the average home tower income of each individual on the average income of the other individuals who are at the same daytime tower. We measure exposure as the individual deviations in absolute terms from this regression line, i.e. the extent to which the income differential between one's own income and other's deviates from the expected level. Specifically, we use the log of the absolute value of the residuals \hat{e}_i as our measure of exposure. To estimate the modifying role of exposure to other strata, we estimate the modified model

$$\text{Events}_{ij} = \alpha + \beta |\ln \bar{y}_i - \ln \bar{y}_j| + \gamma \ln |\hat{e}_i| \times \ln |\ln \bar{y}_i - \ln \bar{y}_j| + \theta' z_{ij} + v_{ij}$$

yielding a margin effect of an increase in the log income difference of $\beta + \gamma \ln |\hat{e}_i|$.

Measuring clustering. Clustering in different networks is computed as the share of triplets of nodes that are closed (i.e. conditional on A communicating with B and C, B and C also communicate with each other). The data are weighted by communication intensity assigning each triplet a weight proportional to the arithmetic mean of the intensity going out, estimated with the `tnet` package in R³⁵.

Results

Income differences and communication intensity. Our main findings are presented in Fig. 1. In Fig. 1a, we plot the average income rank of recipient towers against the income percentile of the senders' towers. We notice a strong tendency of over-proportional in-group communication, especially among the highest income groups. Except for the extremes of the distribution, the relationship is monotonic.

For instance, the expected income of the communication partner increases by about 15% of average income when moving from the 1st to the 70th percentile in the income distribution; the effect is comparable when moving from the 70th to the 100th percentile. The variation in sender tower income ranks is higher than the variation in the average receiver rank due to regression toward the mean effects—if communication is to some extent random, senders at the top and bottom percentiles have their average receiver income rank “biased” towards the sample mean. Still, this is a clear indication that the link strength between two nodes in the network is inversely related to their income difference.

Figure 1b quantifies the relationship between communication and income differences, while controlling for spatial distance and a range of other control variables (see Table A.2 in Supplementary Material for the corresponding regression table). A 10% increase in the income difference between two towers is correlated with 2.98 fewer communication events. This is a substantial number, as less than 43% of tower pairs have 3 or more communication events. To assure that the relationship is not driven by communication within family, we exclude communication occurring within the same night time tower from the analysis. However, this communication represents less than a tenth of a percent of events and less than half a percent of the amount of communication, so the effect is negligible. In the Supplementary Material we also report on how the degree of segregation varies across age and gender (Fig. A.4).

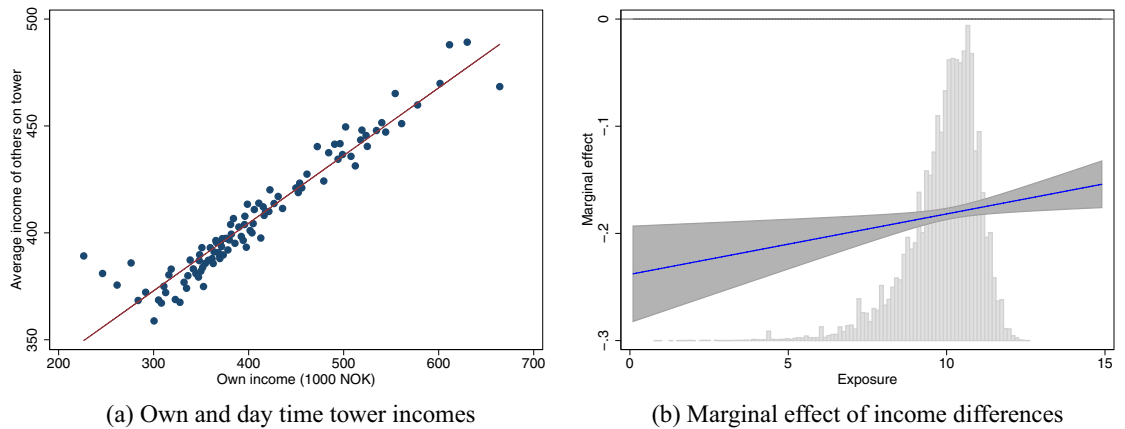


Figure 2. Daytime exposure to other social strata and segregation. Panel (a) shows the relationship between own income and average income at the daytime tower. Exposure to other socioeconomic groups is measured as the absolute deviation from this regression line. Panel (b) shows the distribution of exposure and the marginal effect of income differences on total communication intensity as a function of exposure.

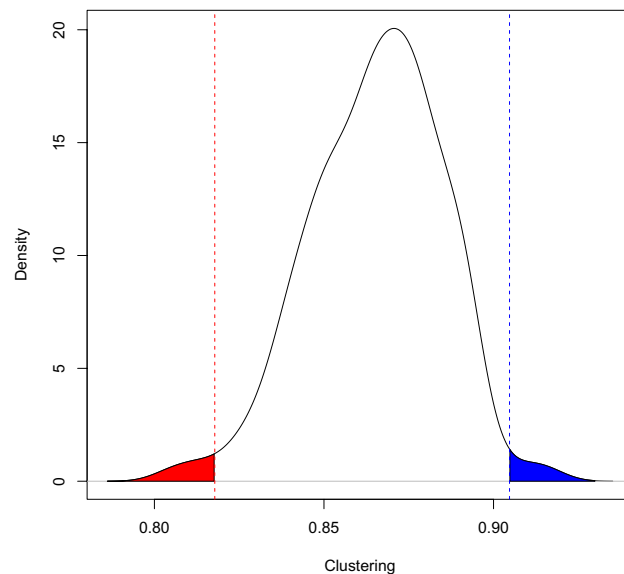


Figure 3. Clustering in networks. The figure shows the distribution of the communication weighted clustering coefficient, in random samples of 100 towers, as well as the clustering coefficient observed in the poorest (red line—to the left) and richest (blue line—to the right) 100 towers.

Daytime exposure to other social strata. In this section, we investigate the relationship between exposure to other socioeconomic groups and segregation. We find a strong positive relationship between own income and other's income ($R^2 = 0.8$), illustrated in Fig. 2a. Hence, spatial income segregation in society is upheld during daytime as well.

Figure 2b reveals a negative correlation of between tower income differentials for all levels of exposure. The positive slope shows that the extent of segregation is decreasing in exposure to other socioeconomic strata (see Table A.3 in Supplementary Material for corresponding regression table). We notice that the total number of events as well as probability of communicating with a tower is less correlated with income differentials for individuals who experience stronger exposure to other income groups during the day as the interaction term is positive. Still, this difference is relatively small in magnitude and there is a negative correlation with income differentials for all groups.

Clustering within income groups. To study the details of the clustering patterns, we compare the level of clustering among the rich and among the poor. The weighted clustering coefficient among the richest and poorest 100 towers is 0.90 and 0.82, respectively. To put the numbers in perspective, we compute the same coefficient for random draws of 100 towers. In Fig. 3, we show the distribution of the simulated groups as well as the two

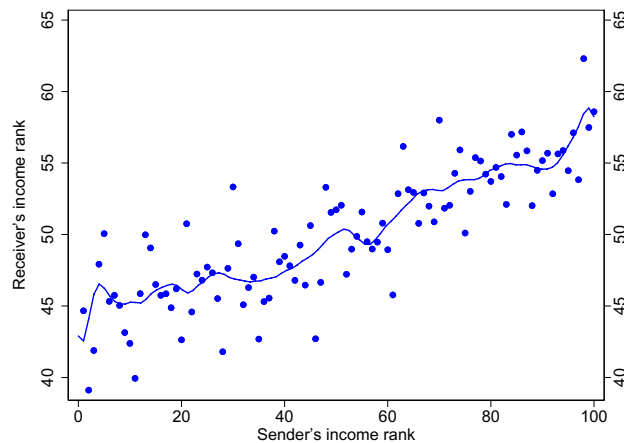


Figure 4. Segregation in the Asian data: average income of sender and receiver. The figure corresponds to Fig. 1a. The figure shows the average income percentile of receivers by the income percentile of the sender as well as local linear smoothing of the relationship. Within tower communication is disregarded.

realized values. Only 1.6% of the simulations are below the value for poorest towers whereas only 1.4% are above the level of the richest towers, indicating very different networks in the two subsamples. In the Supplementary Material we also report the weighted clustering coefficients when using the 25, 50, 100 and 200 poorest and richest towers—the results remain highly stable (Table A.4 in Supplementary Material).

Correlations in a large South Asian city. To verify the external validity of our findings, we undertake the same analysis as in the Norwegian data with the data from a South Asian city. The city is one of the largest in Asia, outside China. Figure 4 replicates Fig. 1a for the Asian data with 111 million links (see Table A.6 in Supplementary Material for corresponding regression table). We find a strikingly similar pattern. Compared to the results for Oslo, it appears that the social segregation among the top 5% is even stronger, and quite extreme compared the rest of the sample. Although the effects of income differences on communication intensity is evident, it is somewhat weaker, possibly because cell phone ownership is less widespread among the poor. This is partially an artifact of the change in the distribution of income between the two countries, but comparing normalized beta coefficients we see that there is a change in effect size as well.

Discussion

Social contact across groups is important for society since it generates trust and collaboration^{8,9}. It is also important at the individual level since networks influence economic opportunities such as finding jobs⁴. Measuring the extent of social contact is difficult as it requires data on the patterns of interpersonal contact in a population. The combination of mobile phone data and income register data carries promise with regards to monitoring the extent and patterns of social segregation.

Our results demonstrate that social segregation is strong in Oslo, in the sense that there is an over-proportional share of communication within income groups. This implies that people from different income groups have less contact with each other. We also examine the extent to which exposure to other socioeconomic groups, measured as neighbourhood during work hours, modifies the social segregation. Social segregation remains strong, even after controlling for daytime exposure to other social strata. Moreover, the degree of social clustering is particularly high among citizens in rich neighborhoods. We also find similar social segregation patterns in a large south-east Asian city.

There are several limitations in this study. First, our data does not allow for any causal interpretations. For instance, in discussing the relationship between exposure to other social strata and the extent to which an individual is socially segregated there may be self-selection whereby more socially integrated individuals choose to work in more socially integrated areas. Second, our measure of daytime exposure to other social strata has an obvious weakness: even though an area is socially diverse, it does not with necessity imply that people interact across socioeconomic groups. Third, whereas income data from Oslo is based on complete registers, the study from a large South East Asian city relies on self reported income. This limits comparability, since self report may be biased and data is likely not missing at random. At the same time, the analysis based on income ranks is less sensitive to this potential weakness.

Future studies should further aim at measuring social segregation over time, something we cannot do with our data due to anonymization. Such a study could perhaps also lead to a better estimate of the causal relationship between exposure and segregation.

Data availability

The data that support the findings of this study are available from two different third parties: Statistics Norway and the telecommunications company. Restrictions apply to the availability of these data, which were used under

license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the third parties. The name of the telecommunications company is available from the authors upon request.

Received: 4 November 2021; Accepted: 17 March 2022

Published online: 19 April 2022

References

1. Jackson, M. O. Networks in the understanding of economic behaviors. *J. Econ. Perspect.* **28**, 3–22 (2014).
2. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444 (2001).
3. Gelman, R., Loewenstein, G., Moene, K. O. & Zari, L. The preference for belief consonance. *J. Econ. Perspect.* **30**, 165–88 (2016).
4. Girard, Y., Hett, F. & Schunk, D. How individual characteristics shape the structure of social networks. *J. Econ. Behav. Organ.* **115**, 197–216 (2015).
5. Allport, G. W. *The Nature of Prejudice* (Basic books, 1954).
6. Pettigrew, T. F. Intergroup contact theory. *Annu. Rev. Psychol.* **49**, 65–85 (1998).
7. Pettigrew, T. F. & Tropp, L. R. A meta-analytic test of intergroup contact theory. *J. Pers. Soc. Psychol.* **90**, 751 (2006).
8. Fukuyama, F. *Trust: The Social Virtues and the Creation of Prosperity* (The Free Press, 1995).
9. Putnam, R. D. *Bowling Alone: The Collapse and Revival of American Community* (Simon and Schuster, 2001).
10. Gambetta, D. *Trust: Making and Breaking Cooperative Relations* (Blackwell, 1988).
11. Hofstra, B., Corten, R., van Tubergen, F. & Ellison, N. B. Sources of segregation in social networks: A novel approach using facebook. *Am. Sociol. Rev.* **82**, 625–656 (2017).
12. DiPrete, T. A., Gelman, A., McCormick, T., Teitler, J. & Zheng, T. Segregation in social networks based on acquaintanceship and trust. *Am. J. Sociol.* **116**, 1234–83 (2011).
13. Bailey, M., Cao, R., Kuchler, T., Stroebel, J. & Wong, A. *Measuring Social Connectedness* Working Paper 23608 (National Bureau of Economic Research, July 2017). <https://doi.org/10.3386/w23608>. <http://www.nber.org/papers/w23608>.
14. Wessel, T. Does diversity in urban space enhance intergroup contact and tolerance?. *Geogr. Annal. Ser. B Human Geography* **20**, 5–17 (2009).
15. Musterd, S., Marcińczak, S., van Ham, M. & Tammaru, T. Socioeconomic segregation in European capital cities. Increasing separation between poor and rich. *Urban Geography* **38**, 1062–1083 (2017).
16. Kalmijn, M. Inter-marriage and homogamy: Causes, patterns, trends. *Annu. Rev. Sociol.* **24**, 395–421 (1998).
17. Currarini, S., Jackson, M. O. & Pin, P. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica* **77**, 1003–1045 (2009).
18. Currarini, S., Jackson, M. O. & Pin, P. Identifying the roles of race-based choice and chance in high school friendship network formation. *Proc. Natl. Acad. Sci.* **107**, 4857–4861 (2010).
19. Blumenstock, J. & Fratamico, L. Social and spatial ethnic segregation: A framework for analyzing segregation with large-scale spatial network data. In *Proceedings of the 4th Annual Symposium on Computing for Development*, 11 (2013).
20. Echenique, F. & Fryer, R. G. Jr. A measure of segregation based on social interactions. *Q. J. Econ.* **122**, 441–485 (2007).
21. Finseraas, H., Hanson, T., Johnsen, Å. A., Kotsadam, A. & Torsvik, G. Trust, ethnic diversity, and personal contact: Experimental field evidence. *J. Public Econ.* **20**, 20 (2019).
22. McLaren, L. M. Anti-immigrant prejudice in Europe: Contact, threat perception, and preferences for the exclusion of migrants. *Soc. Forces* **81**, 909–936 (2003).
23. Blumenstock, J., Cadamuro, G. & On, R. Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076 (2015).
24. Onnela, J.-P. et al. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci.* **104**, 7332–7336 (2007).
25. Onnela, J.-P., Arbesman, S., González, M. C., Barabási, A.-L. & Christakis, N. A. Geographic constraints on social network groups. *PLoS One* **6**, e16939 (2011).
26. Palchykov, V., Kaski, K., Kertész, J., Barabási, A.-L. & Dunbar, R. I. Sex differences in intimate relationships. *Sci. Rep.* **2**, 20 (2012).
27. Kovanen, L., Kaski, K., Kertész, J. & Saramäki, J. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proc. Natl. Acad. Sci.* **110**, 18070–18075 (2013).
28. Toomet, O., Van Der Leij, M. & Rolfe, M. Social networks and labor market inequality between ethnicities and races. *Netw. Sci.* **1**, 321–352 (2013).
29. Toomet, O., Silm, S., Saluveer, E., Ahas, R. & Tammaru, T. Where do ethno-linguistic groups meet? How copresence during free-time is related to copresence at home and at work. *PLoS One* **10**, e0126093 (2015).
30. Xu, Y., Belyi, A., Santi, P. & Ratti, C. Quantifying segregation in an integrated urban physical-social space. *J. R. Soc. Interface* <https://doi.org/10.1098/rsif.2019.0536> (2019).
31. Telenor Research. *Telenor Research Digital Frontrunners Survey 2014* (Telenor Research, 2014).
32. Green, P. & Sibson, R. Computing Dirichlet tessellations in the plane. *Comput. J.* <https://doi.org/10.1093/comjnl/21.2.168> (1978).
33. Robinson, W. S. Ecological correlations and the behavior of individuals. *Am. Sociol. Rev.* **15**, 351–357 (1950) (ISSN: 00031224).
34. Kish, L. A procedure for objective respondent selection within the household. *J. Am. Stat. Assoc.* **44**, 380–387 (1949).
35. Opsahl, T. & Panzarasa, P. Clustering in weighted networks. *Soc. Netw.* **31**, 155–163 (2009) (ISSN: 0378-8733).

Author contributions

A.K., J.T.L. and B.-A.R. devised the research design assisted by J.B. and P.R.S. J.B. and P.R.S. prepared the data for analysis, and J.T.L. analyzed the data with input from A.K. and B.-A.R. J.T.L. and P.R.S. visualized the data with input from A.K. and B.-A.R. A.K., J.T.L. and B.-A.R. wrote the manuscript with inputs from J.B. and P.R.S.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-10273-1>.

Correspondence and requests for materials should be addressed to B.-A.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022