


# SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

## Interaction data from the Copenhagen Networks Study

Piotr Sapiezynski<sup>1</sup>, Arkadiusz Stopczynski<sup>1</sup>, David Dreyer Lassen<sup>2</sup> & Sune Lehmann<sup>1,2\*</sup> 

We describe the multi-layer temporal network which connects a population of more than 700 university students over a period of four weeks. The dataset was collected via smartphones as part of the *Copenhagen Networks Study*. We include the network of physical proximity among the participants (estimated via Bluetooth signal strength), the network of phone calls (start time, duration, no content), the network of text messages (time of message, no content), and information about Facebook friendships. Thus, we provide multiple types of communication networks expressed in a single, large population with high temporal resolution, and over a period of multiple weeks, a fact which makes the dataset shared here unique. We expect that reuse of this dataset will allow researchers to make progress on the analysis and modeling of human social networks.

### Background & Summary

The purpose of collecting the Copenhagen Networks Study (CNS) dataset was to accelerate our understanding of social systems. In particular, we were interested in the following major topics: Measuring networks across modes of communication; Modeling temporal social networks; Modeling spreading processes on social networks; Analyzing and modeling human mobility; Understanding the interplay between mobility and social behavior; Privacy.

Because of our focus on understanding social networks, we enrolled a group of participants (more than 700 freshmen at the Technical University of Denmark) likely to constitute a highly interconnected network. Due to the scale of the study, the amount of raw data collected was substantial: each participant uploaded between 50–100 megabytes of data per day, resulting in new data per day in the range of 50 to 100 gigabytes.

Here, we cannot share the entire raw dataset, below we motivate our choice of which selection of data to publish. The privacy of the study participants is central in the Copenhagen Networks Study, as documented through the study design<sup>1</sup>, as well as our work on privacy discussed in the next section. In a complex dataset, such as ours, it is virtually impossible to provide guarantees regarding re-identification of users while preserving its value for the stated research purposes<sup>2</sup>. In preparing the data for publication, it was therefore necessary to restrict the data released as described below to make de-anonymization as difficult as possible, but without compromising the dataset's usefulness for research. In addition to obstructing de-anonymization, each step listed below serves the second purpose of limiting the potential harm to data subjects in the unlikely case of a successful re-identification attack.

- (a) *Limiting the types of data available.* As part of Copenhagen Networks Study we collected information beyond the dataset we make available here, such as location traces and WiFi logs. Such data-types carry high risk of re-identification through publicly available information. In Denmark, for example, the physical address of every citizen (and phone number) is by default public and published in an open index. Further, geo-located tweets or Instagram posts, social physical activity app data, etc can also be used to re-identify geospatial data. To avoid this attack, we limit our release to information that cannot be easily cross-correlated with public datasets.
- (b) *Limiting the timespan of released information.* The data released only have relative time-stamps. This makes it difficult to cross-correlate the released data with external information. Releasing a full year of data would make it trivial to identify holidays and reconstruct the absolute timestamps. Thus, we do not publish the absolute starting time of the data, but note that the dataset starts on a Sunday during school term.
- (c) *Delaying the data release.* In the European Union, phone metadata retention is limited to two years. By waiting beyond the retention period we limit the probability that a person with access to the network operator

<sup>1</sup>DTU Compute, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark. <sup>2</sup>Center for Social Data Science, DK-1353, Copenhagen, Denmark. \*email: [sljo@dtu.dk](mailto:sljo@dtu.dk)

data (CDR, call detail records) can use such data to re-identify individuals in our dataset. A similar threat is why we do not release Facebook activity data, which could also be used to re-identify individuals via data from inside Facebook (which retains data indefinitely). Furthermore, at the same time, in the case of re-identification, older data is less harmful as it is less likely to be a precise reflection of data subjects' current social networks and behaviors, and even more so since students have moved on from university. In terms of research, this information from a few years ago is just as useful as at the time of collection.

Because the data was recent at the time of publishing most of our research, we took a cautious approach and did not release it then. We are only able to release it now, after a careful balancing of threats and research usefulness.

These considerations on privacy versus types of data released also impact our view on reproducibility given this dataset. The aim of this data release is first and foremost to enable use of rich multi-layer network data for new work, while still respecting participant privacy. That being said, however, much of the work published so far from this dataset was similarly based on four-week subsets (which might or might not overlap with the provided time period). Thus, the timespan of the data is not a limitation to replicability of already published work.

In addition to the network data we release in this paper, the overall CNS study collected detailed high-resolution GPS location (sampled every 5 minutes), information on nearby WiFi routers and cellular towers, screen on/off status, battery charge level, as well as demographic and questionnaire information on all participants<sup>1</sup>. The questions in the 2013 deployment included The Big Five Inventory<sup>3</sup>, Rosenberg Self Esteem Scale<sup>4</sup>, Narcissism NAR-Q<sup>5</sup>, Satisfaction With Life Scale<sup>6</sup>, Rotter's Locus of Control Scale<sup>7</sup>, UCLA Loneliness scale<sup>8</sup>, Self-efficacy<sup>9</sup>, Cohen's perceived stress scale<sup>10</sup>, Major Depression Inventory<sup>11</sup>, The Copenhagen Social Relation Questionnaire<sup>12</sup>, and Panas<sup>13</sup>, as well as health- and behavior-related questions.

A dataset describing human behavior with the richness captured in the CNS study, inevitably raises questions of privacy and personal data. In the CNS data collection, privacy was therefore not only important for the sake of participants, but also an active area of research. In collecting the data, we also had to answer the question, 'how can we work on these data while respecting the privacy of the study participants?'. Therefore, we now briefly discuss overall privacy concerns and challenges. The research project and data collection was registered with and approved by the Danish Data Supervision Authority before data collection commenced. All data was collected with informed consent and with every participant able to withdraw from the study and have their data deleted. This protocol, implemented in 2012 and 2013, was in effect similar to the rules being introduced with the EU General Data Protection Regulation (GDPR) which came into effect in May 2018. In the present release, to comply with GDPR, the data has been stripped of personally identifying information and the data has been reduced in such a way that there is no reasonable likelihood of re-identification occurring.

Details on the actual implementation and broader philosophy of ensuring privacy in sensor-driven human data collection can be found in<sup>1</sup> and<sup>14</sup>. Here, we will remark on two integrated components of the privacy strategy. First, it is well known that formal informed consent can be insufficient to meet actual privacy demands as construed by participants<sup>15</sup>. To address this, we, in addition to the written informed consent paragraph, conducted numerous presentations of the project to students before they signed up, published blog posts, and answered questions using Facebook. Second, we designed a 'quantified self' module allowing participants to access and visualize their own – and only their own – data traces<sup>16</sup>. As the students interacted with these tools, they were able to develop a better understanding of the nature and the depth of the collected data, thus making their consent more informed, or – as was the case for a single student – choose to withdraw from participation.

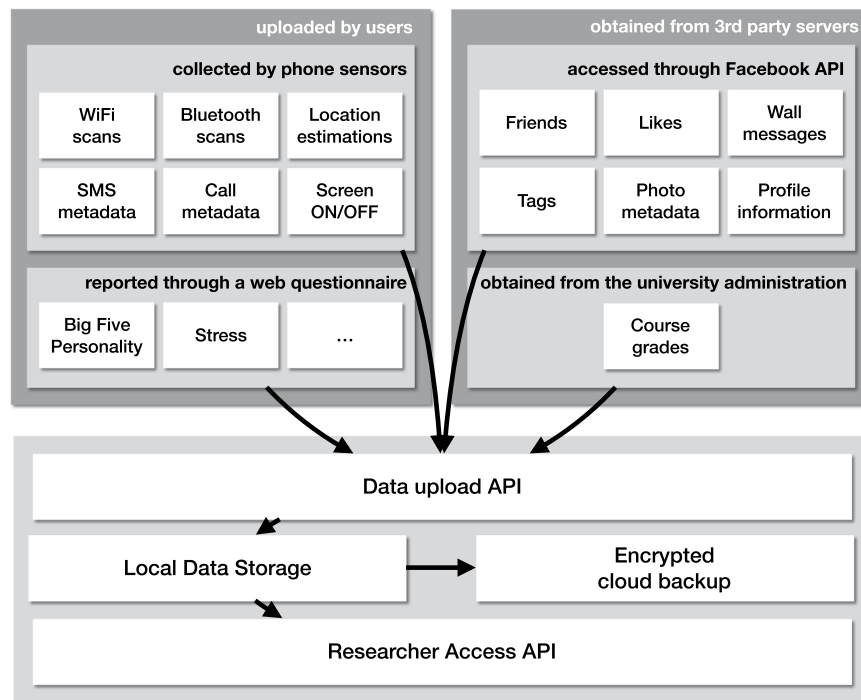
Our work on collecting data from smartphones does not stand alone. In the section below, we provide a brief overview of related work. Mobile phones have been a source of data on human activity and interaction since the early 2000s. They have been used to collect data broadly (coarse grained, sampled data describing millions of individuals) or deeply (fine grained data on fewer individuals). On the large scale, teams have studied the connections between individuals on a societal level in Belgium and Great Britain<sup>17,18</sup>, as well as the mobility of millions of individuals<sup>19,20</sup>. At the other end of the spectrum, teams from MIT's MediaLab have pioneered highly detailed studies of smaller populations. The landmark study is the *Reality Mining dataset*<sup>21</sup>, but more recently many updated studies have been published, in part run by teams at MIT<sup>22–25</sup> as well as the Nokia Research Center in Lausanne<sup>26</sup> and Aalto University<sup>27</sup>. Other similar technologies for measuring social networks have also been developed, for example based on RFID tags, and provide an important alternative to cellphones as social sensors<sup>28–31</sup>. In terms of size, CNS increased the number of participants by almost a full order of magnitude compared to state-of-the-art high-resolution studies<sup>21,25,26</sup>.

Finally, we address the dataset's potential for reuse. The dataset collected as part of CNS has already been used for research in a number of areas. There is a number of publications covering technical aspects of data collection and analysis<sup>16,32–35</sup>. Another set of papers focused on modeling and analyzing network structure<sup>36–41</sup>, epidemiology<sup>42–44</sup>, as well as work on human mobility<sup>45–50</sup>, and privacy<sup>14,46</sup>. Additionally, there is a body of work that goes beyond the stated goals of the CNS project. For example, researchers studied behavioral differences between the two sexes<sup>51,52</sup>, along with studies on academic performance<sup>52–54</sup>, activity patterns<sup>55</sup>, sleep patterns<sup>56–58</sup>, and much more.

The broad and varied research that has already been published based on this dataset underscores its richness. Given that we are only able to release network data, we expect reuse of this dataset to focus on the modeling and analysis of multi-layer temporal networks and we hope that the data released here will allow researchers to make progress on understanding human social networks.

## Methods

The Copenhagen Networks Study accumulated data from a number of channels: smartphones, online questionnaires, and third parties. The data collection system was designed to ensure privacy of the participants and maintain access control to the data, and is described in detail in Stopczynski *et al.*<sup>1</sup>. Figure 1 presents a simplified



**Fig. 1** A schematic view of the data collection for CNS. See main text for a detailed description and Stopczynski *et al.*<sup>1</sup> for a full overview.

overview of the system. Data from all sources were collected first on a central on-premises server. The data were then stripped of personally identifiable information and replicated both locally and as an encrypted cloud backup. Pseudonymized data were then made available to approved researchers via access-controlled API. In this work we make a subset of this dataset available to the public for the first time<sup>59</sup>.

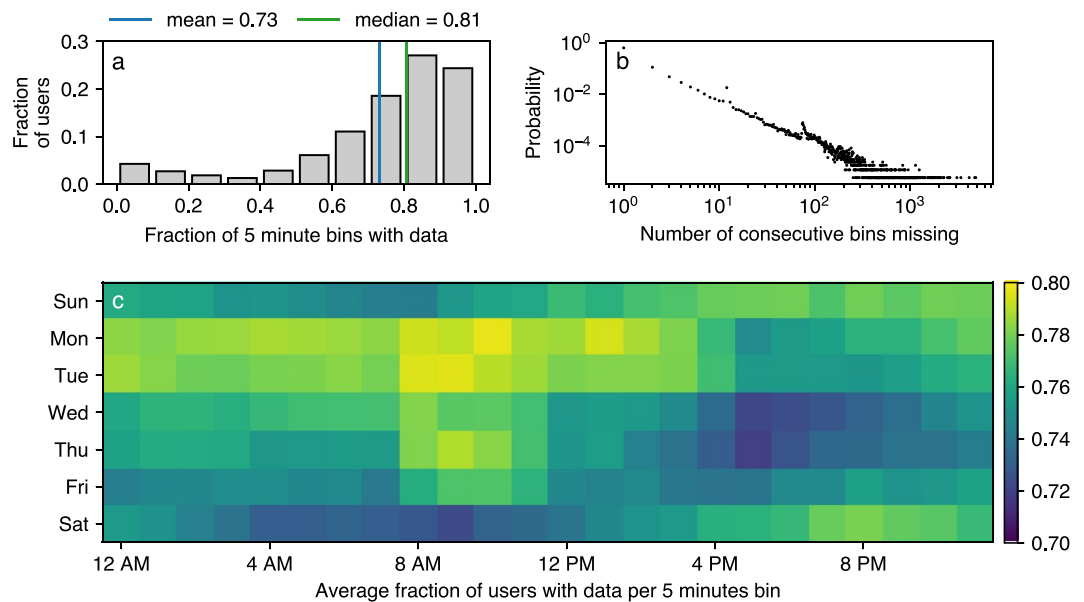
In the sections below we explain the details of the collection methods for each channel.

**Smartphone data collection.** Each participant in the study was equipped with an Android-based Google Nexus 4 smartphone and required to install the data collection software from Google Play Store. Participants agreed to use this study issued device as their primary phone. The data collection software was based on the Funf platform (<http://www.funf.org/>); its source code along with all the modifications we introduced, is open and available on Github (<https://github.com/OpenSensing/funf-v3>). The software triggered data collection from multiple channels: cell phone location (from AGPS), nearby cell towers, WiFi routers, and Bluetooth devices, as well as collected this information whenever another application on the phone requested it from the system. Additionally, each day we collected the meta-data logs of calls and short messages. Our previous work provides a full description of the collected data<sup>1</sup>, here we focus on the three channels which we now make available to the public: Bluetooth data as well as call and SMS metadata. In terms of node-metadata, we include the gender of each participant.

**Bluetooth data.** Bluetooth is a wireless communication standard designed to provide connectivity over distances of up to 10 m (30 ft). Each device in the experiment was configured to be *discoverable* at all times, and to *discover* nearby Bluetooth devices every five minutes. During the discovery process (or *scanning*), a device sends probe requests and receives responses from all nearby Bluetooth discoverable devices. Each response contains a unique identifier of the discoverable device, which also uniquely corresponds to the study participant carrying the device, enabling us to track proximity events between the study participants. Additionally, the device measures and reports the Received Signal Strength (RSSI), which can be (roughly) mapped to physical distance: a high RSSI means that the two devices are physically close, a low measure indicates that they are further apart or that there are obstacles inbetween. In previous work we investigated the interplay between the distance and RSSI in detail<sup>32</sup>.

To prepare the data for release we followed the same pre-processing steps as in other work we published based on this dataset (e.g.<sup>36,42</sup>):

1. We removed identifiers belonging to discovered devices that were not in the experiment, and mapped the participating devices' identifiers to their users.
2. We quantized the time of each scan into bins of five minutes.
3. Within each timebin we found all instances of users  $A$  and  $B$  discovering each other, reported the one with the highest RSSI, and discard others.
4. The information of directionality (whether user $_A$  discovered user $_B$  or vice versa) is discarded.
5. In bins where user $_A$  was actively scanning, but found no other Bluetooth devices in proximity, we reported the alter ID as  $-1$  and the received signal strength as 0.



**Fig. 2** Statistics on data quality. (a) Missing data per user. The data availability – measured as the fraction of 5 minute timebins in which data is available – varies across users. Half of the users have at least 81% of data available. (b) Distribution of time-bins with missing data. Most commonly only a few bins are missing. The visible peak at 12 bins, corresponds to 1 hour of data—the interval at which the phones moved collected data into encrypted files—and could be caused by file corruption on the device. (c) Data availability across the week. The data availability is the highest during working hours, when the majority of interactions occur.

6. In bins where user<sub>A</sub> discovered other Bluetooth devices but not other study participants, we reported the alter ID as  $-2$  and the highest received signal strength measured. We do not report the type of the discovered device.

The data is presented as a temporal, weighted edge list, and each edge is described using (1) the timestamp of the beginning of the timebin in seconds (because of the quantization of time into five minute bins the timestamp is reported in the multiples of 300 seconds), (2, 3) the IDs of users who discovered each other, (4) the measured received signal strength. Note, that in some of the published work (e.g.<sup>36</sup>), we performed the additional step of triadic closure, i.e. if user<sub>A</sub> discovered user<sub>B</sub> and user<sub>B</sub> discovered user<sub>C</sub>, we assumed proximity between user<sub>A</sub> and user<sub>C</sub> regardless of whether they discovered one another. Since there is no meaningful RSSI to assume in such cases, we do not perform this step here and instead we leave it to the researchers using this data to decide which approach is appropriate for their specific analysis.

**Calls and short messages.** Call and message logs were obtained from the smartphones every day. For privacy reasons, we did not capture or store the content of the interactions, only the metadata. Since the participants were required to use the provided smartphones as their primary phones and to reveal their phone number, we matched the entries in the call logs to the participants' identities. Each record in the call logs is in the form of timestamp, user<sub>A</sub>, user<sub>B</sub>, and call duration (in seconds). Each record in the SMS logs is in the form of timestamp, user<sub>A</sub>, and user<sub>B</sub>. In both cases the data is organized such that user<sub>A</sub> initiates the interaction and user<sub>B</sub> is the recipient.

**Facebook data.** Most of the participants of the study voluntarily opted in to authorize data collection from Facebook. We used the official Facebook API and the access tokens provided by participants to collect their Facebook data every day. The data we collected include all the participants' activity, characteristics, and the contents of their News Feed. Here, we release a static snapshot of the friendship network among the participants at the end of the observation period (links to non-participants are removed). This friendship network is presented in the form of a static edge list.

**Data quality.** In this section we report on the quality (availability) of the Bluetooth data. Phones in the experiment were set to scan for Bluetooth every five minutes. We therefore divide the four weeks observation period into 8064 five minute periods ( $4 \text{ (weeks)} \times 7 \text{ (days)} \times 24 \text{ (hours)} \times 12 \text{ (five minute periods)} = 8064$ ). The data quality (availability) of each user is the fraction of these periods in which they have scan data (there were actively scanning and/or scanned by another user). Figure 2 summarizes the Bluetooth data availability concerns. Panel a displays the distribution of data quality as defined above: the median availability is 0.81, meaning that half the users have data in 81% of timebins or more. In Panel b, we show that in most cases only few consecutive bins are missing, with a small peak at one hour. Panel c emphasizes the fact that missing data is correlated in time. The best quality is observed during working hours and the worst on the night between Friday and Saturday. We expect higher data coverage when more users are interacting – even if one user's phone fails to report scan results, that

column name	column description
timestamp	Timestamp in seconds from the beginning of the observation period (as reported by the device. Note, that because of differences in the internal clock of different devices, some of the measurements will not be perfectly aligned.)
user_a	ID of one user (ego).
user_b	ID of the other user (alter). 0–850 for participants of the study, –1 for empty scans, –2 for any non-participating device.
rssi	Received Signal Strength Indication, measured in dBm, a rough proxy for distance between devices (the higher the absolute value, the higher the distance)
<b>Summary:</b> 5,474,289 records, 706 total users	

**Table 1.** Bluetooth interactions. These are listed in `bt.csv` and are formatted as described above.

column name	column description
timestamp	Timestamp in seconds from the beginning of the observation period
user_a	ID of the user initiating the call
user_b	ID of the other user receiving the call.
duration	Duration of the interaction in seconds, or –1 for missed calls
<b>Summary:</b> 3,600 records, 540 total users	

**Table 2.** Phone calls. These are stored in `calls.csv` and are formatted as described above.

column name	column description
timestamp	Timestamp in seconds from the beginning of the observation period
user_a	ID of the user sending the message.
user_b	ID of the user receiving the message (non-participants were removed).
<b>Summary:</b> 24,333 records, 577 total users	

**Table 3.** SMS data. Short messages are listed in `sms.csv` are formatted as described above.

column name	column description
user_a	ID of user A
user_b	ID of user B
<b>Summary:</b> 6,429 edges, 800 total users	

**Table 4.** Facebook friendships. The static snapshot of Facebook friendships is stored `fb_friends.csv`, and formatted according to the description in this table.

user is likely to be scanned by others around them. The effect of the Friday night missing data is a combination of fewer study participants nearby each other and the study participants neglecting to charge their phones during a night out.

## Data Records

All data is available in Figshare<sup>59</sup>. A description of the Bluetooth interaction data is available in Table 1, call info is listed in Table 2, text message data descriptions are in Table 3, the Facebook data description is in Table 4, and description of the gender information is stored in Table 5.

## Technical Validation

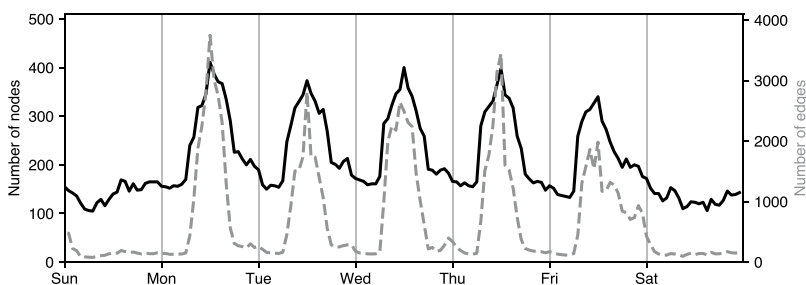
In this section we describe and visualize the properties of the networks, providing readers with an overview which we hope will facilitate working with the data.

**Person to person proximity (Bluetooth data).** *Temporal dynamics.* The properties of the presented Bluetooth data reflect the highly dynamic and circadian nature of interactions between the participants of the study. Figure 3 summarizes 168 hours (one week) of data by reporting the number of active participants (i.e. participants in the vicinity of other participants) as well as the number of active edges in the network. Note that a small part of the network is active over night and during weekends: some students share accommodations or have dorm rooms adjacent to one another. During daytime on weekdays, the network grows and becomes much more connected. Notice as well, that the overall properties of the network during Friday are different than other days of the week, with interactions continuing late into night hours.

*Temporal aggregates.* Given the volume and high temporal resolution of the Bluetooth data, one might introduce temporal aggregation to simplify the analysis of the data. It is, however, important to note that the structure of the network changes drastically as the aggregation window grows. Figures 4 and 5 illustrate the effects of

column name	column description
user	ID of user
gender	gender of user; 0 for male, 1 for female
<b>Summary:</b> 788 total users	

**Table 5.** Binary descriptors of participants' gender. These are recorded in `genders.csv`, and formatted as explained above.



**Fig. 3** Bluetooth activity reveals a clear circadian and weekly rhythm. Black solid line shows the number of active nodes, while the dashed gray bars show the number of active edges.

aggregation. At five minutes—the underlying temporal resolution of the data—participants form multiple small, disjoint groups. Figure 4 highlights how the biggest connected component in the graph grows with aggregation up to 40 minutes. Figure 5 shows the effect of aggregating the data for up to 24 hours: the structure of the network is not clear from the network layout, and the average degree grows from  $\sim 2$  to  $\sim 30$ . For more details on how aggregation changes the structure of the network and the subsequent implications for epidemic modelling, refer to Stopczynski *et al.*<sup>43</sup>. For a more detailed description of the network structures at the highest resolution, and the implications for modeling of social interactions, refer to Sekara *et al.*<sup>36</sup>.

Figure 6 shows that the aggregates of the Bluetooth networks are very dense: in the weekly aggregates between 10% and 20% of possible links are active at least once, and during the entire observation 30% of all possible links were active.

**Telecommunication network.** Figure 7 shows that the participants of the study prefer text messages to making phone calls, with an order of magnitude higher number of SMSes than calls. The telecommunication networks appear to be complementary to the person-to-person network: participants resort to text messages and phone calls in more the evenings and weekends, when there are fewer proximity events. As shown in Fig. 6, the difference in the density of aggregate networks is not pronounced as strongly as in the sheer volume of communications. There is a positive correlation between the number of messages and phone calls dyads exchange.

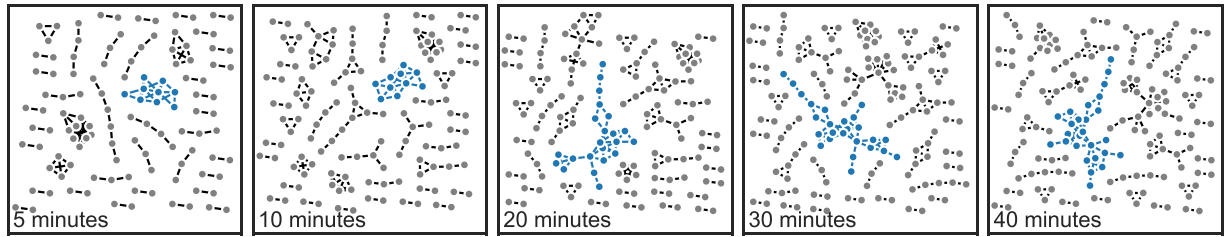
**Network comparisons.** The overlap between networks is an interesting avenue of investigation, which we explored to some extent in our previous work<sup>32,38,40</sup>. Figure 8 shows the overlap between most active dyads in different networks. We see that over 80% of short messages are exchanged among 15% of dyads that have the most physical proximity (panel A), and that 89% of dyads call each other are also in physical proximity at least once during the observation (panel B).

**Data loading.** The data is released as CSV files. We show how data can be loaded using the Pandas<sup>60</sup> package in Python. All data files are directly loadable with a basic call of `pandas.read_csv()`.

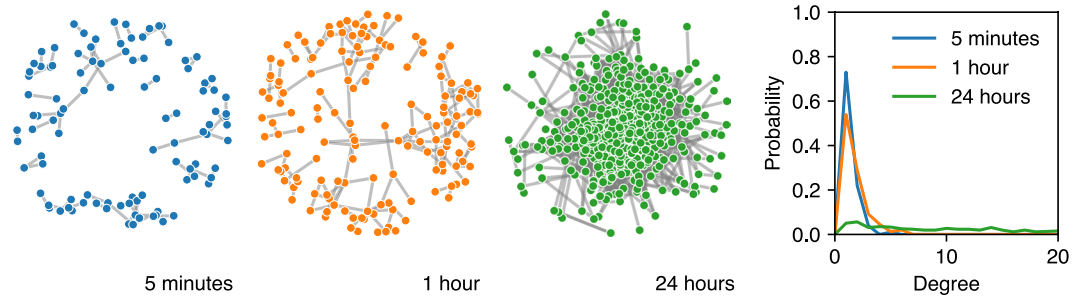
**Bluetooth network visualization.** The iPython notebook released with the data shows a visualization (using NetworkX<sup>61</sup> and Matplotlib<sup>62</sup>) of the temporal Bluetooth network by considering the network structure at a single 5-minute bin, similar to the visualization in Fig. 5. At this high temporal resolution, the network consists of many small connected components which can be directly used for network analysis<sup>36</sup>.

We note that while the network is preprocessed to be symmetric wrt. RSSI values (we store the higher value of RSSI between two users), the components in the network are not necessarily fully connected: if user<sub>A</sub> saw user<sub>B</sub> and user<sub>C</sub>, but user<sub>B</sub> did not see user<sub>C</sub> (nor vice versa) we do not create a link between user<sub>B</sub> and user<sub>C</sub>. Due to high temporal and spatial resolution of the provided data, users of the of data may consider treating the components as fully connected at given time slice, expecting that spurious connections disappear in analysis over longer periods, see, for example, Sekara *et al.*<sup>36</sup>.

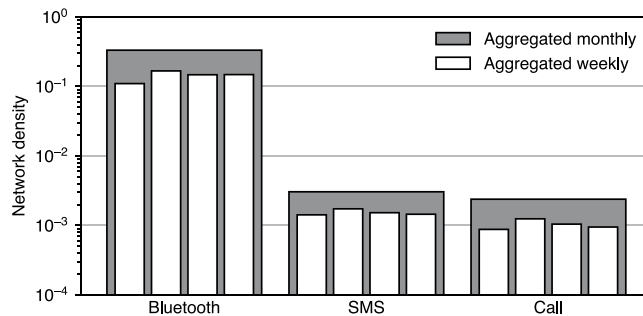
**SMS communication visualization.** The notebook also shows the principle of joining different types of data by visualizing the number of text messages sent between users of different genders in the study. Using a consistent user id in all the data types, allows for straightforward merging of different subsets, allowing us to, for example, consider dynamics of communication separately for different genders.



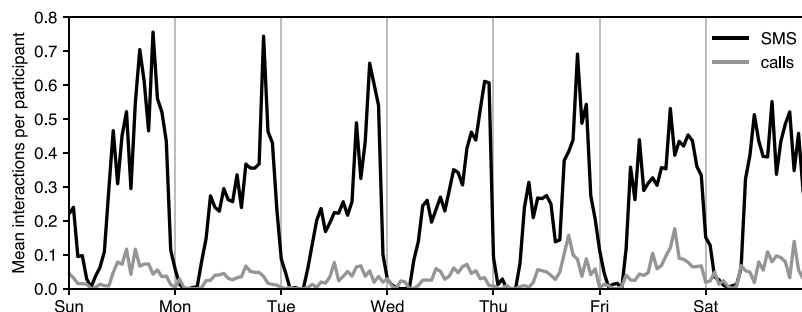
**Fig. 4** Temporal aggregation of the Bluetooth network. The biggest connected component (shown in blue) grows steadily as we increase the duration of the aggregation window from five minutes—the underlying sampling frequency in the data—to 40 minutes.



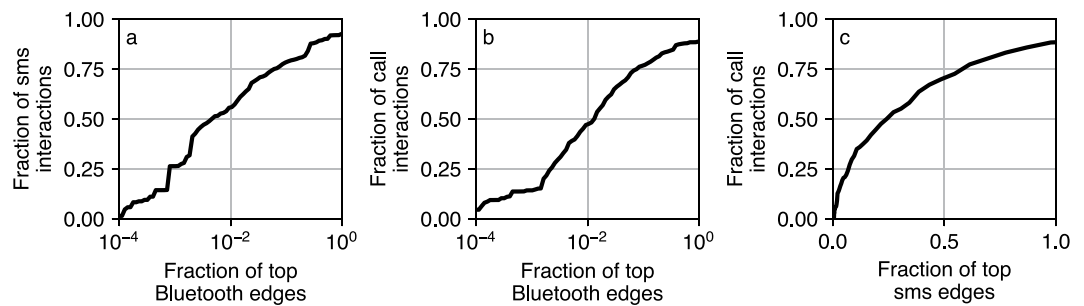
**Fig. 5** Temporal aggregation of the Bluetooth network. By using daily aggregates important structures are obscured<sup>36</sup>.



**Fig. 6** Relative network density in weekly and monthly aggregates. Density is the fraction of dyads that are active vs the number of possible dyads. In the nested bar-charts, the dark gray bars show data aggregated by month, while light bars show data aggregated weekly.



**Fig. 7** Telecommunications activity. Contrary to interactions in the physical space, the participants call/text each other most during evenings and weekends. The black solid line shows the mean number of text messages per participant as a function of time, whereas the gray line shows the corresponding mean number of phone calls.



**Fig. 8** Overlap between most active dyads in different networks. Dyads who interact a lot in physical space also communicate most through calls and sms. 1% of most active Bluetooth dyads correspond to (a) 56% short messages exchanged and (b) 47% of phone calls. (c) 1% of most active sms dyads correspond to 6% of phone calls.

### Code availability

Alongside the data, we provide an iPython notebook showing basic data loading and use (see `Copenhagen_Networks_Study_Notebook.ipynb` in the Figshare data repository<sup>59</sup>). The notebook is intended to showcase the basic approaches to working with the released data.

Received: 13 March 2019; Accepted: 21 November 2019;

Published online: 11 December 2019

### References

1. Stopczynski, A. *et al.* Measuring large-scale social networks with high resolution. *PLOS One* **9**, e95978 (2014).
2. Rocher, L., Hendrickx, J. M. & de Montjoye, Y.-A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* **10**, 3069 (2019).
3. John, O. P. & Srivastava, S. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* **2**, 102–138 (1999).
4. Rosenberg, M. *Society and the adolescent self-image (rev.)*. (Wesleyan University Press, 1989).
5. Back, M. D. *et al.* Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. *J. Pers. Soc. Psychol.* **105**, 1013 (2013).
6. Diener, E., Emmons, R. A., Larsen, R. J. & Griffin, S. The satisfaction with life scale. *J. Pers.* **49**, 71–75 (1985).
7. Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied* **80**, 1 (1966).
8. Russell, D. W. UCLA loneliness scale (version 3): Reliability, validity, and factor structure. *J. Pers.* **66**, 20–40 (1996).
9. Sherer, M. *et al.* The self-efficacy scale: Construction and validation. *Psychol. Rep.* **51**, 663–671 (1982).
10. Cohen, S., Kamarck, T. & Mermelstein, R. A global measure of perceived stress. *J. Health. Soc. Behav.* 385–396 (1983).
11. Bech, P., Rasmussen, N.-A., Olsen, L. R., Noerholm, V. & Abildgaard, W. The sensitivity and specificity of the major depression inventory, using the present state examination as the index of diagnostic validity. *J. Affect. Disord.* **66**, 159–164 (2001).
12. Lund, R. *et al.* Content validity and reliability of the Copenhagen Social Relations Questionnaire. *J. Aging. Health.* **26**, 128–150 (2014).
13. Watson, D., Clark, L. A. & Tellegen, A. Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* **54**, 1063 (1988).
14. Stopczynski, A., Pietri, R., Pentland, A. S., Lazer, D. & Lehmann, S. Privacy in sensor-driven human data collection: A guide for practitioners. *arXiv* **1403**, 5299 (2014).
15. Strandburg, K. J. *et al.* Privacy, big data, and the public good: frameworks for engagement (2014).
16. Cuttone, A., Lehmann, S. & Larsen, J. E. A mobile personal informatics system with interactive visualizations of mobility and social interactions. In *Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia*, 27–30 (ACM, 2013).
17. Blondel, V., Guillaume, J., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**, P10008 (2008).
18. Eagle, N., Macy, M. & Claxton, R. Network diversity and economic development. *Science* **328**, 1029–1031 (2010).
19. Gonzalez, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 479 (2008).
20. Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).
21. Eagle, N., Pentland, A. & Lazer, D. Inferring social network structure using mobile phone data. *PNAS* **106**, 15274–15278 (2007).
22. Madan, A., Moturu, S., Lazer, D. & Pentland, A. Social sensing: obesity, unhealthy eating and exercise in face-to-face networks. In *Wireless Health 2010*, WH'10, 104–110 (ACM, New York, NY, USA, 2010).
23. Madan, A., Cebrián, M., Lazer, D. & Pentland, A. Social sensing for epidemiological behavior change. In *UbiComp'10*, 291–300 (2010).
24. Madan, A., Farrahi, K., Gatica-Perez, D. & Pentland, A. Pervasive sensing to model political opinions in face-to-face networks. *Perv. Comp.* 214–231 (2011).
25. Aharoni, N., Pan, W., Ip, C., Khayal, I. & Pentland, A. Social fmri: Investigating and shaping social mechanisms in the real world. *Perv. and Mobile Comp* (2011).
26. Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D. & Laurila, J. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin* (2010).
27. Karikoski, J. & Nelimarkka, M. Measuring social relations: Case otasizzle. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, 257–263 (IEEE, 2010).
28. Cattuto, C. *et al.* Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLOS One* **5**, e11596 (2010).
29. Stehlé, J. *et al.* High-resolution measurements of face-to-face contact patterns in a primary school. *PLOS One* **6**, e23176 (2011).
30. Barrat, A. *et al.* Empirical temporal networks of face-to-face human interactions. *Eur. Phys. J. ST.* **222**, 1295–1309 (2013).



31. Panisson, A., Gauvin, L., Barrat, A. & Cattuto, C. Fingerprinting temporal networks of close-range human proximity. In *Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2013 IEEE International Conference on, 261–266 (IEEE, 2013).
32. Sekara, V. & Lehmann, S. The strength of friendship ties in proximity sensor data. *PLOS One* **9**, e100915 (2014).
33. Sapiezynski, P., Gatej, R., Mislove, A. & Lehmann, S. Opportunities and challenges in crowdsourced wardriving. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, 267–273 (ACM, 2015).
34. Wind, D. K., Sapiezynski, P., Furman, M. A. & Lehmann, S. Inferring stop-locations from wifi. *PLOS One* **11**, e0149105 (2016).
35. Sapiezynski, P., Stopczynski, A., Wind, D. K., Leskovec, J. & Lehmann, S. Inferring person-to-person proximity using wifi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1**, 24 (2017).
36. Sekara, V., Stopczynski, A. & Lehmann, S. Fundamental structures of dynamic social networks. *PNAS* **113**, 9977–9982 (2016).
37. Mollgaard, A. *et al.* Measure of node similarity in multilayer networks. *PLOS One* **11**, e0157436 (2016).
38. Mones, E., Stopczynski, A. & Lehmann, S. Contact activity and dynamics of the social core. *EPJ Data Sci.* **6**, 6 (2017).
39. Aslak, U., Rosvall, M. & Lehmann, S. Constrained information flows in temporal networks reveal intermittent communities. *arXiv:1711.07649* (2017).
40. Sapiezynski, P., Stopczynski, A., Wind, D. K., Leskovec, J. & Lehmann, S. Online behaviors of offline friends. *arXiv: 2462825* (2018).
41. Dissing, A. S., Lakon, C. M., Gerds, T. A., Rod, N. H. & Lund, R. Measuring social integration and tie strength with smartphone and survey data. *PLOS One* **13**, 1–14 (2018).
42. Stopczynski, A., Pentland, A. S. & Lehmann, S. How physical proximity shapes complex social networks. *Sci. Rep.* **8** (2018).
43. Stopczynski, A., Sapiezynski, P., Pentland, A. S. & Lehmann, S. Temporal fidelity in dynamic social networks. *Eur. Phys. Jour. B* **88**, 249 (2015).
44. Mones, E. *et al.* Optimizing targeted vaccination across cyber–physical networks: an empirically based mathematical simulation study. *J. Royal Soc. Interface* **15**, 20170783 (2018).
45. Cuttone, A., Lehmann, S. & Larsen, J. E. Inferring human mobility from sparse low accuracy mobile sensing data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 995–1004 (ACM, 2014).
46. Sapiezynski, P., Stopczynski, A., Gatej, R. & Lehmann, S. Tracking human mobility using wifi signals. *PLOS One* **10**, e0130824 (2015).
47. Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S. & Baronchelli, A. Evidence for a conserved quantity in human mobility. *Nat. Hum. Behav.* **1** (2018).
48. Alessandretti, L., Sapiezynski, P., Lehmann, S. & Baronchelli, A. Multi-scale spatio-temporal analysis of human mobility. *PLOS One* **12**, e0171686 (2017).
49. Cuttone, A., Lehmann, S. & González, M. C. Understanding predictability and exploration in human mobility. *EPJ Data Sci.* **7**, 2 (2018).
50. Alessandretti, L., Lehmann, S. & Baronchelli, A. Understanding the interplay between social and spatial behaviour. *EPJ Data Sci.* **7**, 36 (2018).
51. Psylla, L., Sapiezynski, P., Mones, E. & Lehmann, S. The role of gender in social network organization. *PLOS One* **12**, e0189873 (2017).
52. Sapiezynski, P., Kassarnig, V., Wilson, C., Lehmann, S. & Mislove, A. Academic performance prediction in a gender-imbalanced environment. In *FATREC Workshop on Responsible Recommendation Proceedings* (2017).
53. Kassarnig, V. *et al.* Academic performance and behavioral patterns. *EPJ Data Sci.* **7**, 1–16 (2018).
54. Kassarnig, V., Bjerre-Nielsen, A., Mones, E., Lehmann, S. & Lassen, D. D. Class attendance, peer similarity, and academic performance in a large field study. *PLOS One* **12**, 1–15 (2017).
55. Mollgaard, A., Lehmann, S. & Mathiesen, J. Correlations between human mobility and social interaction reveal general activity patterns. *PLOS One* **12**, 1–16 (2017).
56. Cuttone, A. *et al.* Sensible sleep: a bayesian model for learning sleep patterns from smartphone events. *PLOS One* **12**, e0169901 (2017).
57. Aledavood, T., Lehmann, S. & Saramäki, J. Social network differences of chronotypes identified from mobile phone data. *EPJ Data Sci.* **7**, 46 (2018).
58. Rod, N. H., Dissing, A. S., Clark, A., Gerds, T. A. & Lund, R. Overnight smartphone use: A new public health challenge? A novel study design based on high-resolution smartphone data. *PLOS One* **13**, 1–12 (2018).
59. Sapiezynski, P., Stopczynski, A., Lassen, D. D. & Jørgensen, S. L. The Copenhagen Networks Study interaction data. *figshare*, <https://doi.org/10.6084/m9.figshare.7267433> (2019).
60. McKinney, W. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing* 1–9 (2011).
61. Hagberg, A., Swart, P. & Chult, S. & Exploring, D. network structure, dynamics, and function using networkx. Tech. Rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (2008).
62. Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

## Acknowledgements

We acknowledge support from the Villum Foundation (Young Investigator grant to SL) and the Independent Research Fund Denmark (*Sapere Aude* grant to SL) as well as the UCPH2016 grant Social Fabric (main PI, DDL). We are thankful to the Social Fabric steering group: Anders Blok, Jesper Dammeyer, Rikke Lund, Joachim Mathiesen, Morten Axel Pedersen, Julie Zahle.

## Author contributions

P.S. and S.L. planned the paper. P.S., A.S., D.D. and S.L. designed the anonymized data set. P.S. created the figures and created the data sets. P.S., A.S., D.D. and S.L. wrote the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to S.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2019