Visual Analysis of Social Networks in Space and Time using Smartphone Logs

Aidan Slingsby, Roger Beecham, Jo Wood

giCentre, City University London, Northampton Square, London, EC1V 0HB, UK

Abstract

We designed and applied interactive visualisation techniques for investigating how social networks are embedded in time and space, using data collected from smartphone logs. Our interest in spatial aspects of social networks is that they may reveal associations between participants missed by simply making contact through smartphone devices. Four linked and coordinated views of spatial, temporal, individual and social network aspects of the data, along with demographic and attitudinal variables, helped add context to the behaviours we observed. Using these techniques, we were able to characterise spatial and temporal aspects of participants' social networks and suggest explanations for some of them. This provides some validation of our techniques.

Unexpected deficiencies in the data that became apparent prompted us to evaluate the dataset in more detail. Contrary to what we expected, we found significant gaps in participant records, particularly in terms of location, a poorly connected sample of participants and asymmetries in reciprocal call logs. Although the data captured are of high quality, deficiencies such as these remain and are likely to have a significant impact on interpretations relating to spatial aspects of the social network. We argue that appropriately-designed interactive visualisation techniques – afforded by our flexible prototyping approach – are effective in identifying and characterising data inconsistencies. Such deficiencies are likely to exist in other similar datasets, and although the visual approaches we discuss for identifying data problems may not be scalable, the categories of problems we identify may

Email addresses: a.slingsby@city.ac.uk (Aidan Slingsby), roger.beecham.1@city.ac.uk (Roger Beecham), j.d.wood@city.ac.uk (Jo Wood)

be used to inform attempts to systematically account for errors in larger smartphone datasets.

Keywords:
Big data, human behaviour, spatiotemporal, social networks, visual analysis

---

1. Introduction

We increasingly organise our lives through smartphone devices using instant messaging, voice calls, calendars, reminders, social media and location-based services. As a result, log files from such devices and supporting services provide significant opportunities for social scientists and market researchers to identify and understand human behaviour [13, 7] and their social networks [6]. Our interest lies in where such devices are *spatially* aware, where there is the potential to study the *geography* of people's lives. This is because spatial proximity often reduces the need to make contact through smartphones, so neglecting space may neglect important characteristics of associations between participants.

Our aim was to explore the extent to which we can characterise spatial and temporal aspects of social networks using exploratory visual analysis. Using call and GPS logs from the Lausanne Data Collection Campaign [16] as part of our entry [22] to Nokia's Mobile Data Challenge [18], we wanted to draw on techniques from visual analytics to explore four research questions (RQs):

- To what extent can smartphone device logs help us to understand social communication behaviour?
- Can exploratory visualisation techniques help us characterise spatial and temporal aspects of participants' social networks?
- Can linking spatial, temporal and call connectivity patterns help us explain how participants construct their social networks?
- Can providing information about participants help us generalise our findings?

Formulating analytical methods to answer such questions is difficult. Exploratory visual analysis can help through its ability to show overview,

zoom, filter and offer details on-demand in response to the analytical process [20, 25, 2] and is widely advocated for studying human behaviour in time and space [9, 15]. Effective exploratory visual analysis requires tools that offer well-designed interactive graphics that allow relevant aspects of data to be normalised and related, in response to analysts' needs.

Rather than using existing visual analysis software, our approach involved iterative prototyping to design appropriate visualisation and interactive techniques whilst at the same time using the prototype to explore the data. This stems from our view that tool development and visual analysis are not separate processes [27]. Although slower (depending on experience and availability of suitable libraries) the advantage is that the focus of analysis can be more flexibly steered towards questions that arise as a result of the exploration. In our case, as it became clear data issues existed that might impact interpretations made, we were able to adapt our techniques to investigate these. Like many other researchers, we had assumed that such issues with the dataset would not be significant due to the large data sample that was collected over many months in which great lengths were taken [16] to ensure data reflected activities of participants well. Our techniques enabled us to study the data in detail and helped us assess their characteristics and suitability for the task at hand. Many non-visual approaches overlook the importance of exploring data in detail. This even applies to some visual analytics approaches, particularly those that reduce the data from the outset through largely automatic sampling, aggregation or clustering. We assert the importance of using visual analysis for detailed data exploration and for evaluating a dataset's characteristics, limitations and suitability. This is especially important for 'reality mining' studies [10], which aim to automatically infer characteristics of participants, with the implication of scaling to a larger population.

We present and discuss details of how we linked spatial, temporal, call connectivity and participant information, our interactive visualisation design and the conclusions these techniques helped us draw. We reflect on the degree to which identified social networks appear to be embedded in space and time. Our contributions are:

- to advocate a prototyping approach that enables the visual analytics design to adapt to the data, support data enrichment (e.g. georeferencing calls) and the changing issues that arise during visual exploration;
- to suggest interactive visualisation methods suitable for studying spa-

tial and temporal aspects of calls between participants;

- to validate the methods and techniques used by describing patterns in the data, suggesting reasons for them, describing characteristics of the dataset and how they might impact on answering our research questions; and
- to discuss deficiencies in such data that might impact on findings relating to spatial aspects of social networks.

2. Related work

As suggested in the introduction, there is a great deal of work that investigates the potential for inferring human characteristics from these types of data. The term 'reality mining' has been used to characterise a burgeoning research area whereby sensor data on human behaviour are collected and analysed [10]. Information generated from mobile phones is a rich and pervasive source of such data. As well as providing commercially useful information on customers' consumption behaviours, such data offer new means of researching how humans interact with each other and their environment [10]. Within data mining and machine learning disciplines, researchers have interrogated attribute-rich data provided by smartphone logs to predict participants' personal information such as ethnicity, age and marital status, evaluating derived models against the stated attributes of research participants [1]. Studying the communication intensity, regularity and temporal tendency of known research participants, Min et al. propose an approach to characterising contact behaviours that are between family, work and more informal social relationships [17]. In a similar study, Do and Gartica-Perez present a model that aims to discover different interaction types based on known participants' proximity, phone call or email network data [8]. The authors identify classes of communication behaviour relating to routine work and leisure activities, as well as more unique events [8].

Whilst these studies, which aim to automatically infer and describe social interactions, are significant, they have yet to fully explore social-spatial relations: the extent to which certain social communication behaviours or social networks are spatially situated. Incorporating GPS or location data into these analyses could enrich and better enable context-specific interactions between participants to be investigated [8]. Kapler and Wright's GeoTime software is a visual analysis tool initially designed in an intelligence analysis

context and which aims to track events, objects, activities and interactions of participants within a combined temporal and geospatial display [14]. The software displays space-time events within a geographic view, with a temporal slider enabling filtering at different temporal resolutions, and further drill down information on specific entities or events available through interaction. The tool enables connected events or entities to be filtered as well as an association analysis, whereby only locations that an individual or set of participants has visited or has been contacted are shown [14]. In addition, MobiVis [19] is a visual analysis tool that uses mobile phone data collected from university staff and students to identify social-spatial information exchanges. Node-link diagrams are used to depict relations between participants, positions and meeting places. They also suggest presenting temporal information in two dimensions: months along one axis, and hours of the day or days of the week along another. This enables the structure of cyclical activities, daily or hourly, to be depicted and analysed in a longitudinal context [19].

Our research intentions are in many ways analogous to those of GeoTime and MobiVis. However, the ultimate objective of GeoTime is to identify individual 'stories' and exceptional activity within large datasets. In our study, as well as identifying specific social-spatial events, in RQ3 and RQ4 we aim to characterise and make more general observations about the nature of social-spatial interactions. Although the MobiVis tool is successful at identifying general patterns of interactions between sets of research participants, we would also argue that the software fails to offer sufficiently rich spatial descriptions. In integrating both social and spatial summaries into a single network graph, the tool discriminates only between discrete spatial categories of places visited. Whilst such an approach may be relevant where only a small set of spatial categories are of interest, since our research ambition is exploratory, we are necessarily interested in exploring social-spatial communication behaviours at multiple spatial scales and contexts.

3. Design

3.1. Design approach

Our approach to visual analysis reflects our view that visualisation design and data analysis are not separate activities [27]. We iteratively design, prototype and test visualisation ideas, generating new research questions that inform the design process. The act of working with the data from the start

by designing and prototyping graphics and interactions to find structure in it gives us a deep understanding of the data that informs our designs. Ideas that do not work can be discarded; those that do are refined to support or address our research questions. We used basic data analysis software for early exploratory ideas and analyses. Processing [12, 11] – a graphically oriented set of Java libraries that facilitate the rapid prototyping of data visualisation designs – was subsequently used. Designs were refined at each iteration, ultimately resulting in the final design we present here. As it became clear that there were issues with data quality, our flexible design approach allowed the focus of our analysis to move towards studying data quality and representation.

Graphically oriented programming frameworks such as Processing provide a large degree of flexibility in design compared to using off-the-shelf tools. A disadvantage is that the extra flexibility afforded inevitably requires more implementation time. However, choosing a framework with which one has experience, using designs and code from previous projects [21] and using third-party libraries [26] helps mitigate this.

3.2. Design

Our tool design was guided by our research questions, for which we needed means to filter, query and view spatial, temporal and participant aspects of the smartphone log data. Crucially, we also needed to be able to view and query connections between participants.

We were fortunate to already have validated techniques and code for studying spatial and temporal aspects of GPS data, giving us a useful starting point. These techniques originally resulted from a user-centred design exercise with animal behaviouralists studying seagull behaviour from birds they tracked with GPS loggers [21, 23]. The resulting tool – developed over an intensive two-week period of workshops, prototyping, feedback and evaluation – allowed the domain specialists to explore their data in a way that was not previously possible using a tool they helped design. The novelty of that work was the close involvement of 'users' and evaluation based on their research questions. We used this validated design and code as a way of exploring space-time patterns of behaviour, modifying the design as necessary for addressing our research questions here, the most significant addition of which related to showing call and text message behaviour.

Figure 1 shows the final design of our tool with three coordinated views. The zoomable map view (Figure 1A) plots the spatial data on a base map

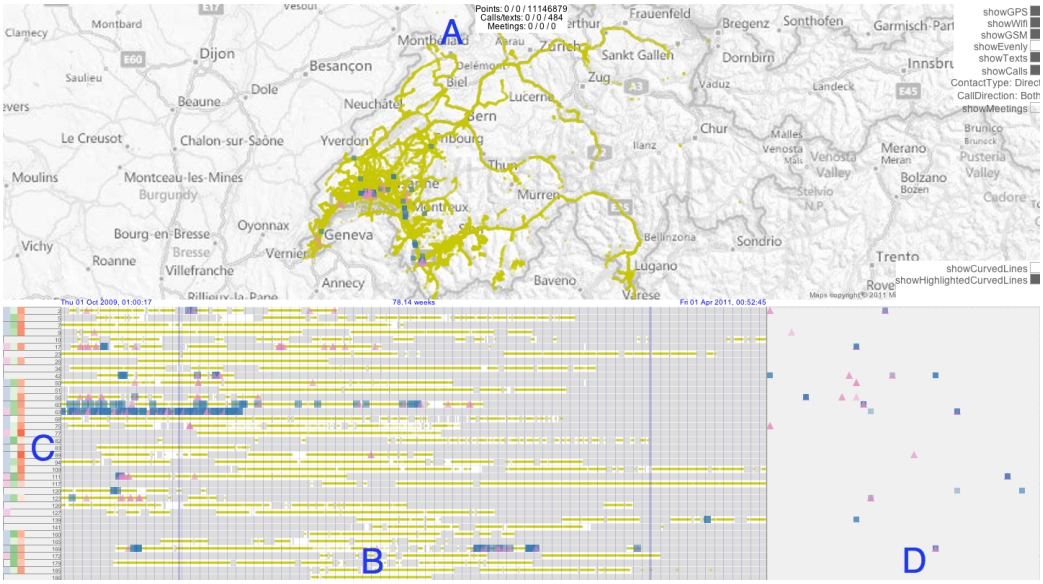

Figure 1: Screenshot of tool, with four coordinated views: [A] Zoomable map that shows all GPS points; blue and purple indicates where calls and texts were made. [B] Zoomable timeline (bottom) with one row per participant showing their GPS positions (yellow), calls (blue) and texts (purple). [C] List of participants where colours indicate gender (blue: male, pink: female), age (dark green: older) and a derived measure of social activity (dark red: more socially active). [D] Matrix (bottom right) of number of calls between participants. See video at `https://vimeo.com/43245266`.

(Bing Maps). The zoomable timeline (Figure 1B) view plots temporal aspects of the data either *(a)* linearly from the start to the end of the study period, *(b)* by day of week or *(c)* by hour of day. Each row of the timeline relates to a participant. On the left, three coloured squares indicate gender (pink=female; blue=male), age (light green=young; dark green=old) and a measure of social activity (light red=low; dark red=high). The matrix view (Figure 1C) show data that relates to pairs of participants, where columns are in the same order as the rows. Each of these views has a 'point mode' (Fig. 1) which shows individual records and a 'density mode' (e.g. Fig. 2). As the name suggests, the 'density mode' directly represents the density of points aggregated to a fixed grid or spatial or temporal unit. The simple density surfaces are computed on the fly by simply counting the number of points within the cells of a grid based on the screen pixels. On the timeline, grid-squares are the same width as the height of the row; on the map, the width and height are 5 pixels – thus, they are dependent on the current zoom

level. Cells are then shaded using an appropriate colour mapping function. We used ColorBrewer [5] Purples' and 'Reds' for density (e.g. Fig. 2) and highlighted density (e.g. Fig. 5) respectively. User-defined maximum colour scaling can be changed interactively and reset to the maximum in view. This interactive colour scaling adjustment helps study the variation in values at different magnitudes. Whilst maps in 'point mode' display individual points and allow each to be queried, the resulting occlusion makes the density of points difficult to determine (try comparing the maps in Fig. 3).

These three views employ coordinated brushing to relate spatial, temporal and participant aspects of the data. Data can be filtered and queried through direct mouse and keyboard interaction with the graphics. Our accompanying video (`https://vimeo.com/43245266`) demonstrates the coordinated brushing [24] and selection across these views that enables locations within a particular temporal window, times within a spatial window and other relationships between data represented by these three coordinated views.

4. Data

Nokia supplied us with 18 months of usage data from 38 participants (personal data were obfuscated and locations around participants' home locations removed [16]) for the Data Challenge. A standard set of log files was available for each participant, relating to different aspects of their smartphone use. We were also supplied with the results of a questionnaire completed by participants and a list of the days in which they participated.

There are some appealing characteristics of these data for answering our research questions. The data were collected over a long period of time. We know some of the characteristics of the participants (through questionnaire responses) allowing us to group and filter participants sharing common characteristics. Like many examples of 'reality mined' data, the data were automatically collected through ubiquitous devices, reducing the chance of gaps caused by people forgetting to turn loggers on or forgetting to charge the battery. There are, nevertheless, some less appealing characteristics. With only 38 participants, we can only suggest techniques through which identified communication behaviours might be generalised, rather than make more concrete claims within this dataset. As the data were collected through ubiquitous devices and efforts to produce good quality data had been made, we were surprised that there were gaps in the data, which we judged as making

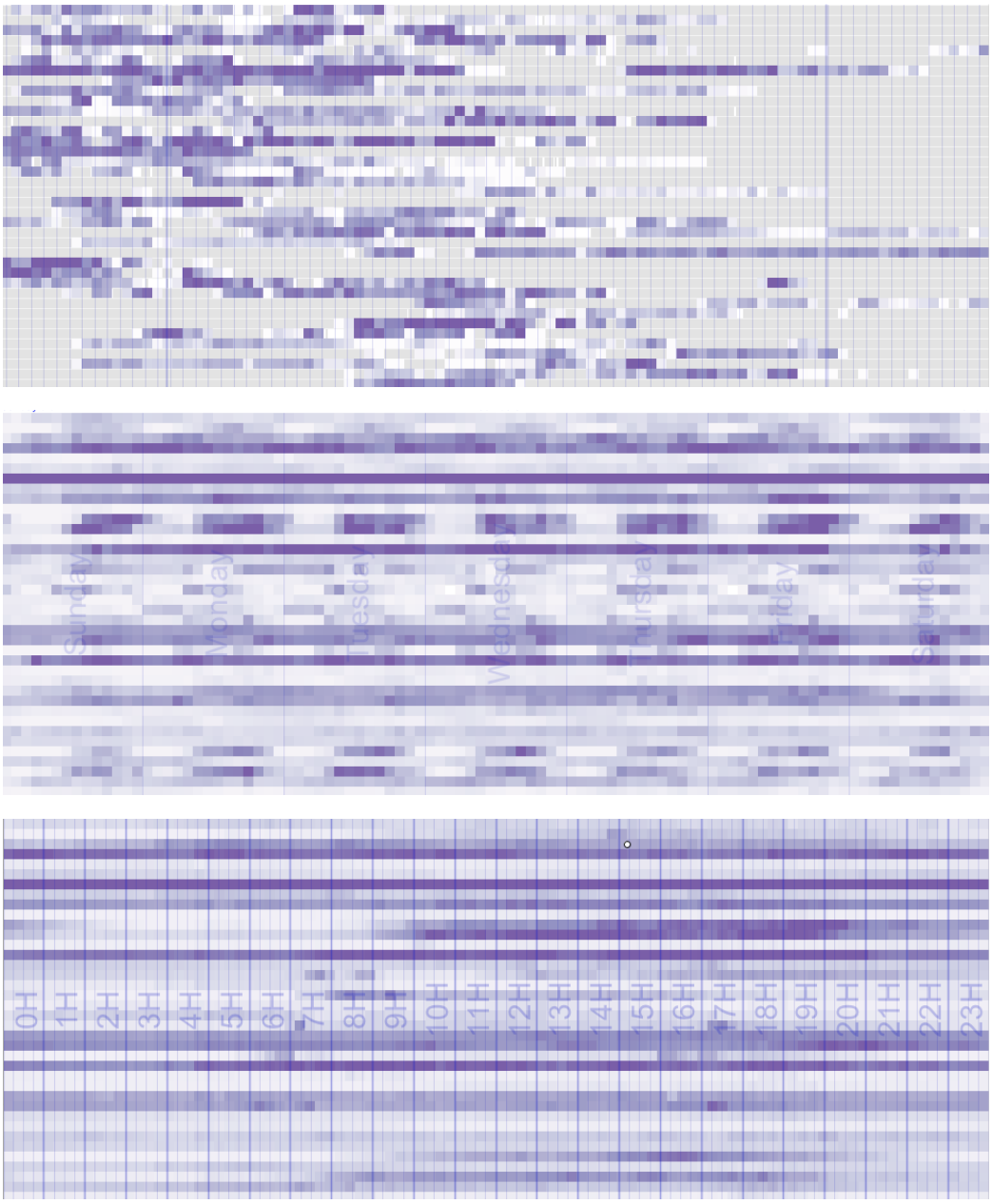

Figure 2: The timeline component of our tool (Fig. 1B) in 'density mode' showing the temporal density of all spatial data (GPS, wi-fi and GSM; 11.7 million records) by user (row) for the whole 18-month period (top; grey indicates when participants did not participate), by day (middle) and by hour (bottom). Data density differences between participants by day and by hour mostly reflect differences in participation duration. Note that the blocky appearance is due to the representation of density being binned into grid cells with the same width as the height – see page 7. Grid lines are displayed when appropriate for the zoom level.

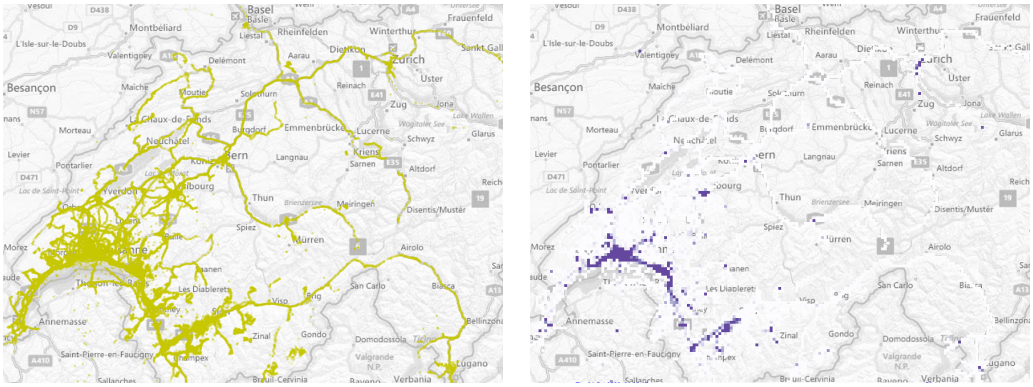

Figure 3: The map component of our tool (Fig. 1A) in 'point mode' (left) and 'density mode' (right), showing spatial aspects of the data in Fig. 2 (all the GPS points, wi-fi positions and GSM records). The 'point mode' shows individual points, allows them to be queried and shows the extent of movement by the participants, but may give a misleading impression of where most participants spend most of their time. The 'density mode' shows that, as expected, participants are in small pockets of areas which tend to correspond to built-up areas.

the data less representative and adversely affecting our ability to answer our research questions.

All the separate log files contained timestamps that enabled them to be related to each other. Some records were continuously and regularly sampled. This included the GPS logs, but network difficulties resulted in gaps, the implications of which we discuss later in this section. Consistent with our approach to design, decisions on how to process the data were informed by its exploration in our tool as it was built and designed.

5. Georeferencing behaviour

Our research questions require data that describe *communication between participants*, and *where* and *when* this communication occurs. Spatial data were supplied as two log files: GPS locations (about every 10 seconds) and estimated locations from wi-fi access points (about every 5 minutes). Additionally, GSM logs provided the identity of GSM base stations in use at about 1 minute intervals. We georeferenced as many of these as we could using GPS and wi-fi data and the resulting set of georeferenced GMS base stations provided a third source of spatial context.

An advantage of having three different sources of spatial data is that

different sources work in different circumstances. GPS works best outdoors with a clear view of the sky, wi-fi works best where there is a high density of wi-fi coverage as well as indoors and GSM works anywhere where there is phone reception. Fig. 2 shows the density of all three sources of data on timelines. Studying Fig. 2, we find there are marked temporal differences between participants and the certainty with which we can locate them. Some only participate for short periods of time; others turn their smartphones off at night.

Fig. 3 shows the spatial distribution of these GPS, wi-fi and GSM data records, with the density map confirming that most data are for Lausanne, along the northern edge of Lake Geneva and Martigny.

5.1. Locating a participant

These three sources of spatial data allowed us to locate an individual at any time. This allowed us to georeference artefacts of participant behaviour including calls and meetings.

GPS generally has a spatial accuracy of less than 10 metres and provides the most accurate position. We know less about the wi-fi positioning because it is based on triangulating from several wi-fi access points whose recorded position may not be accurate. GSM base stations have wider coverage than wi-fi stations but there is usually a higher density in built-up areas. Based on these likely spatial accuracies, we have the highest preference for GPS and the lowest preference for GSM.

To locate the GSM base stations (included in Fig. 3), we find the closest GPS position in time to each GSM record from the same user within a 90 second temporal window. Where there is no GPS position, we do the same for wi-fi positions. Then, for each base station, we find the average position, weighted by the signal strength. We found locations for 64% of GSM base stations (13,789/21,473).

In total, we have 1.6 million GPS positions, 1.8 million wi-fi positions and 7.7 million GSM positions with which we can locate participants.

To locate participants, we used two temporal windows: a smaller one and a larger one. Starting with the smaller temporal window, we identify the closest GPS position in time within this window. If none exists, we do the same for wi-fi position and then GPS. If we have still failed to obtain a position, we do the same with the larger temporal window. The initial smaller temporal window allows for less precise georeferencing if a position available close in time which we consider preferable to a more precise location

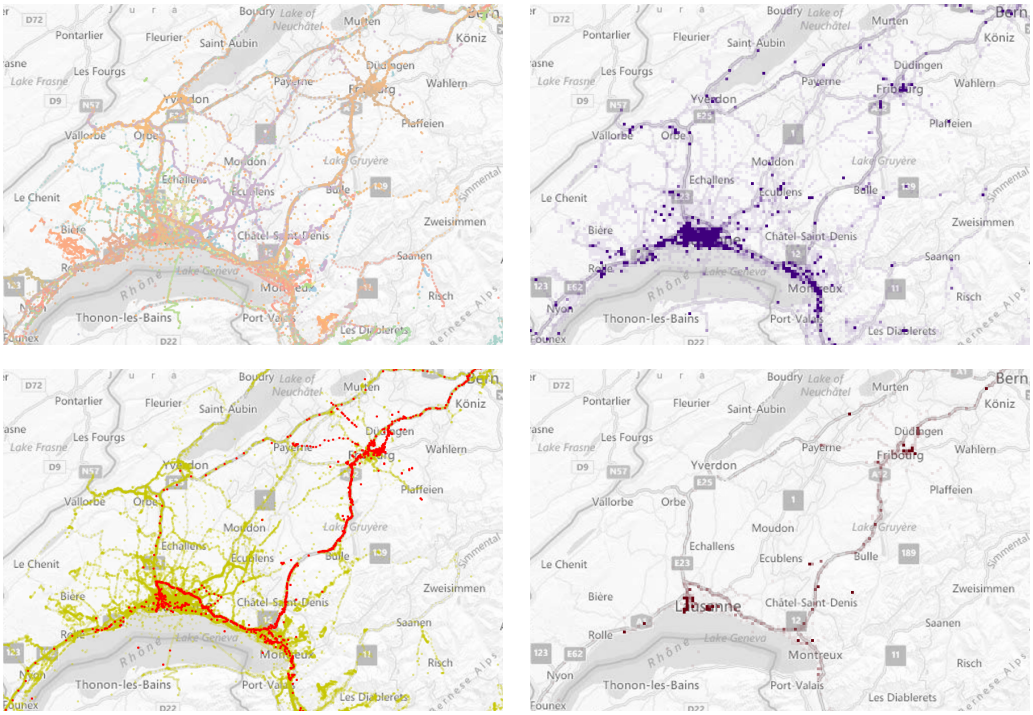

Figure 4: The map component of our tool (Fig. 1A) showing participant locations sampled at 1-minute intervals for days in which they participated over the 18 month period. *Top left:* Map in 'point mode' coloured by participant. Occlusion makes it difficult to see who spent time where. *Top right*: As top left, but in 'density mode', where density indicates the amount of time a participant spent there. *Bottom left:* Participant 139 highlighted in 'point mode' (using interactive brushing) showing the places he has visited. *Bottom right:* Participant 139 highlighted in 'density mode' (using interactive brushing), showing that he spends a significant amount of time in Fribourg.

further away in time. Appropriate temporal window sizes depend on the sampling rate of the three sources of spatial data and assumptions about participants' movement. We chose threshold sizes that were informed by our visual techniques.

5.2. Gaps and temporal sampling in the spatial record

Gaps in the spatial data make the spatial record of participants incomplete. Where gaps correspond to specific *places* (e.g. GPS blackspots), those places and the specific spatial behaviours associated with those places will be underrepresented. If the device is turned off whilst the participant is stationary, we might assume the participant is where they were when the device

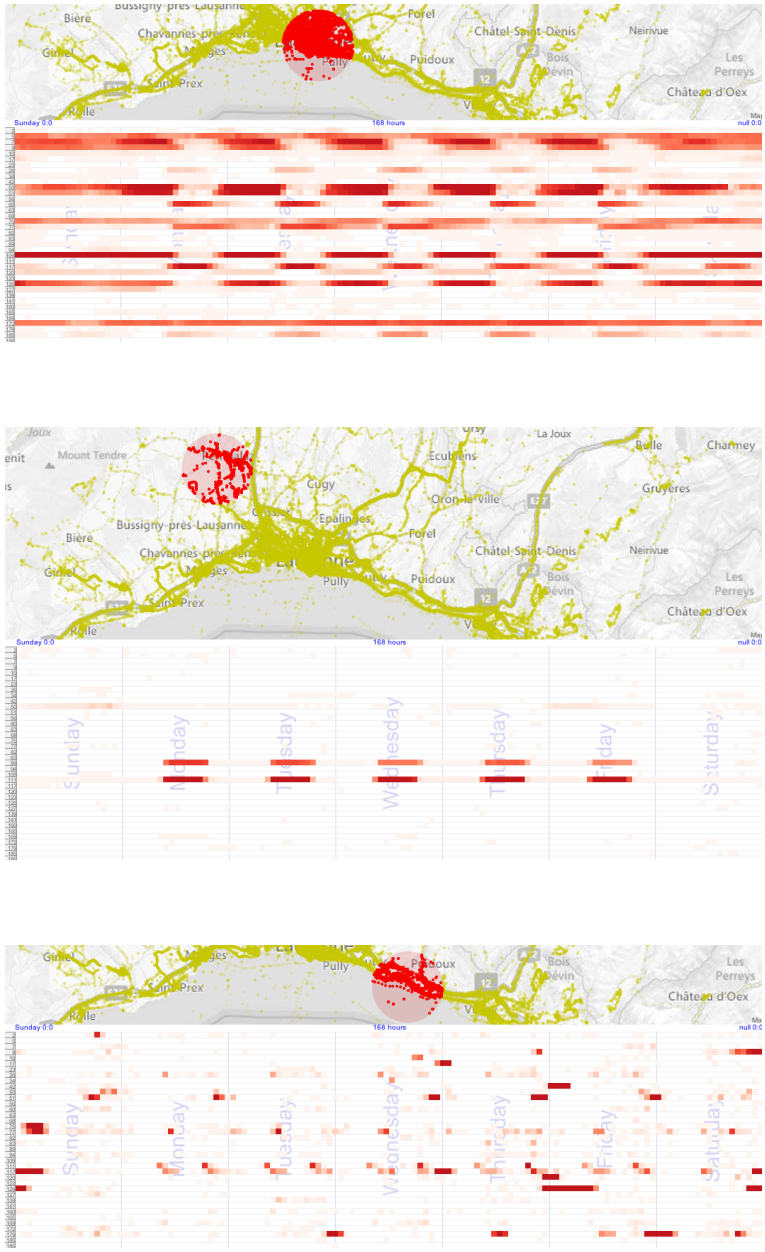

Figure 5: Studying the temporal signature (in 'density mode') of places highlighted with brushing may give clues as to how the space is used. *Top:* The temporal signature suggests the brushed area on the map is a residential address for some participants (night-time only) and a work address for others (day-time only). *Middle:* The temporal signature suggests the brushed (highlighted) place is a workplace for two participants. *Bottom:* The temporal signature suggests the brushed areas on the map is a place of transit through which participants travel.

was in use, but we can only make this assumption if the space-time context of this event is meaningful: if, for example, the device is turned off at an individual's likely home in the evening or workplace in the morning.

Uneven temporal sampling also affects how spatially representative the data are. Use of all three sources of spatial data in Fig. 3, each of which has gaps and different temporal sampling, means that densities do not represent how much participants used space – it only represents the spatial data of our incomplete record of participant behaviour.

To produce a spatial record that more closely reflects participants' use of space, we regularly sampled locations at 1 minute intervals, but only for the days in which individuals participated. Using the method described in section 5.1, we used an initial temporal window of 5 minutes and subsequent temporal window of 1 day. The initial smaller temporal window ensures that even a georeferencing method with low precision can be used if close enough in time; the larger temporal window will ensure that locations are derived even if the participant remained at the same location all day. There is some subjectivity here, but visual analysis helps identify suitable thresholds.

Fig. 4 shows regularly sampled locations for participants where the density surfaces reflect time spent at that location. Interactive brushing can be used to select participants to find out how they use space, as in Fig. 4 (bottom).

5.3. Exploring when and how places are used

Interactive brushing to highlight places on the map or temporal windows on the timeline helps us explore their temporal and spatial signatures, giving us clues about how these places are used. Fig. 5 (top) suggests the area selected is a residential address for some participants, a place of work for others and both for other participants. Fig. 5 (middle) suggests that the highlighted place is a workplace for two participants. Fig. 5 (bottom) appears to be a place of transit through which participants travel.

5.4. Findings associated with location

We found large gaps in the spatial record which vary by user. After regularly sampling locations in time, we obtained a relatively consistent record of where participants have been and for how long. 'Point mode' and 'density mode' maps show how space was used and interactive brushing helps us relate views, showing participants' use of space, when people visited particular

places and the times at which places were used. In Fig. 5, we characterise places based on the temporal signature of when they were used.

6. Call logs

Details of the 81,044 incoming and outgoing calls and text messages received, sent and missed were supplied. Entries contained a timestamp and the sent/received phone number (obfuscated). Using the method outlined in section 5.1 with an initial temporal window of 1 minute and a subsequent window of an hour, we were able to georeference 9% (7,968) of contacts (calls and text messages) with GPS, 13% (10,259) with wi-fi and 46% (37,524) with GSM. We were unable to georeference 32% (25,293) of contacts.

6.1. Spatial and temporal distribution of calls

Fig. 6 shows that, although the spatial distribution of calls and text messages broadly follows that of the spatial distribution of participants, two main centres around Lausanne and around Montreux (to the east) are split by a highly frequented area, but one that is more associated with transit in Fig. 5 (bottom). Fig. 6 also shows that calling behaviour is similar on each day of the week, perhaps suggesting that calling behaviour is not driven by work life. Throughout the day, there is an increase in activity around lunchtime, a slight lull in the late afternoon and then greater use in the evening. As expected, participants do make many calls at night, but two participants make or receive many calls between 0600 and 0700.

6.2. Inspecting individual calls

We can investigate these calls further by using interaction. In Fig. 7 (left), the timeline has been zoomed to 0600-0700 and the mouse pointer has been used to identify records. Seven call records exist under the mouse (as shown by the tooltip). Left/right clicking with the mouse cycles forwards/backwards through these records, and the identified record is shown on the map (if it has been georeferenced) as well as the timeline. In this case, participant 111 was called by someone outside of the study, but who is a contact of participant 139. If we then filter these to only show direct calls between participants, Fig. 7 (right) shows that over the 18 months of the study, these two participants called each other directly (as indicated by the curved line that indicates direction [28]) a number of times, as well as calling others, many of whom are contacts of other participants in the study. Calls

Figure 6: All voice calls and text messages made or received by participants. *Top left:* 'Point mode' maps of the 55,751 voice calls (blue squares) and text messages (pink triangles) that we could georeference. *Top right:* 'Density mode' maps of voice calls and text messages that give a better impression of where most calls are made and text messages received and sent. *Middle:* Calls and texts by day, showing similar daily behaviour. *Bottom:* Calls and texts by hour, showing that more voice calls and text messages are sent and received during the evening and there is a lull in the late afternoon.

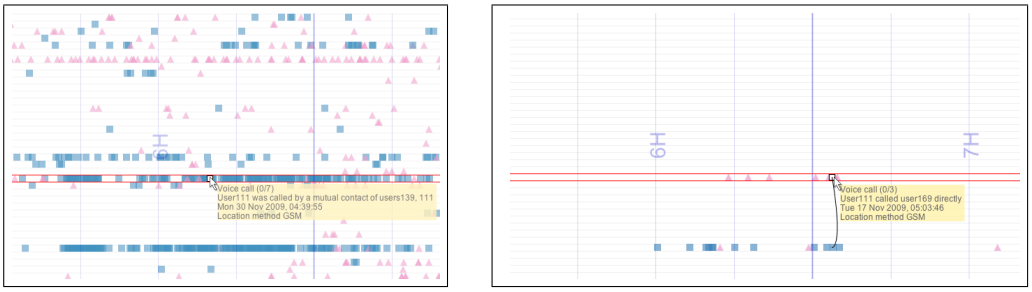

Figure 7: Zooming in on the timeline and inspecting the high density of calls and text messages between 0600 and 0700 observed in Fig. 6. *Left:* The tooltip shows that the call identified by the mouse pointer is to a mutual contact of participants 139 and 111. *Right:* Calls are filtered to only show direct calls to participants. The identified call is a direct call between the two participants.

between these two participants were made at relative spatial proximity. This leads us to speculate that this early morning phone activity is associated with car sharing.

6.3. Direct calls between participants

Fig. 8 shows the 0.6% (484) of calls and text messages were made directly between the 26 participants for whom we knew the telephone numbers. Of these, the location that the calls were received or made could be determined in 57% (270) of cases. We could determine both the receiver and sender in 44% (214) of cases. Fig. 8 shows that one participant dominates voice and text calls. The sparse call matrix suggests that participants are not well connected.

Whilst exploring the nature of direct calls by brushing the map as shown in Fig. 9, we established that the overwhelming majority of direct calls involved one location and two participants, participant 63 and 123. We were surprised that participant 123 had no record of any of these reciprocal calls. Part of the reason is illustrated in Fig. 10: participant 123 was absent during some of the study period. However, many reciprocal calls are also not recorded, even on days in which the individual participated. This is further evidence of gaps within the smartphone logs, and again suggests the importance of more systematic data checking when dealing with such behavioural datasets.

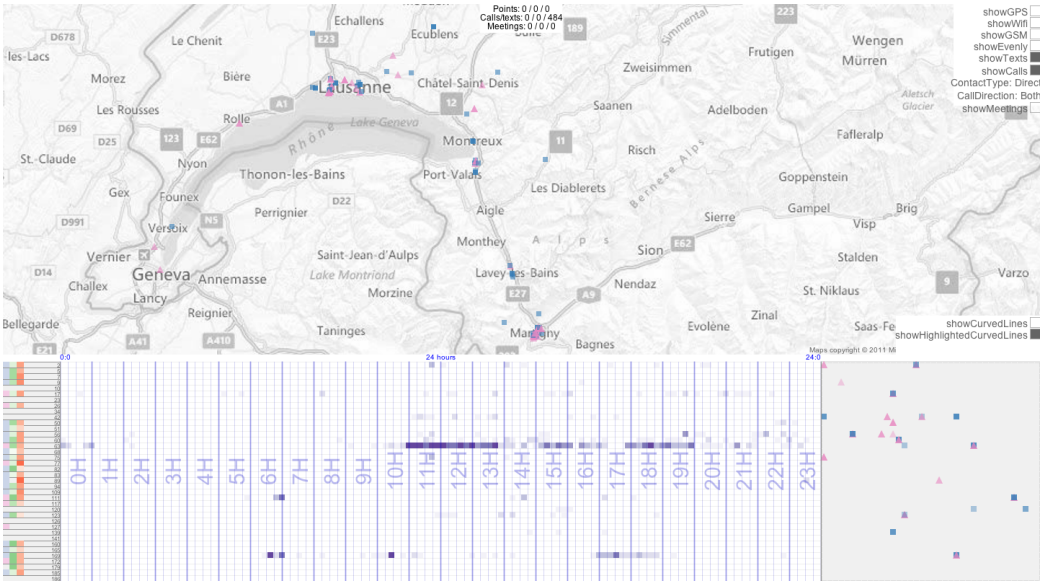

Figure 8: Voice calls and text messages have been filtered to only show the 484 direct calls between participants. Direct calls are dominated by one participant. The call matrix is very sparse, most cells only representing a few calls. These are key pieces of information that help assess how representative the call log is of participants.

6.4. Participant contacts

The sparsity of direct contacts led us to consider other means of describing participants' social network. We had only a subset of participants from the Lausanne Data Collection Campaign, so wondered about the nature of the original social network.

Only 484 direct calls were made between participants. Of the 80,560 direct calls made to people outside the study, 13,786 of these were made to people with at least one contact from within the study. Given how few calls were made between participants of the study, we were surprised that 2,814 calls were made to people outside the study with two contacts within the study, 1,840 calls were made to those with 3 contacts with the study, 1,121 calls were made to those with 4 contacts with the study, 1,106 calls were made to those with 5 contacts with the study, 1,004 calls were made to those with 6 contacts with the study and 871 calls were made to those with 7 contacts with the study. This latter figure is still double the number of direct calls between participants. Studying patterns with 'mutual contacts' outside the study provides useful information about the nature of the social network.

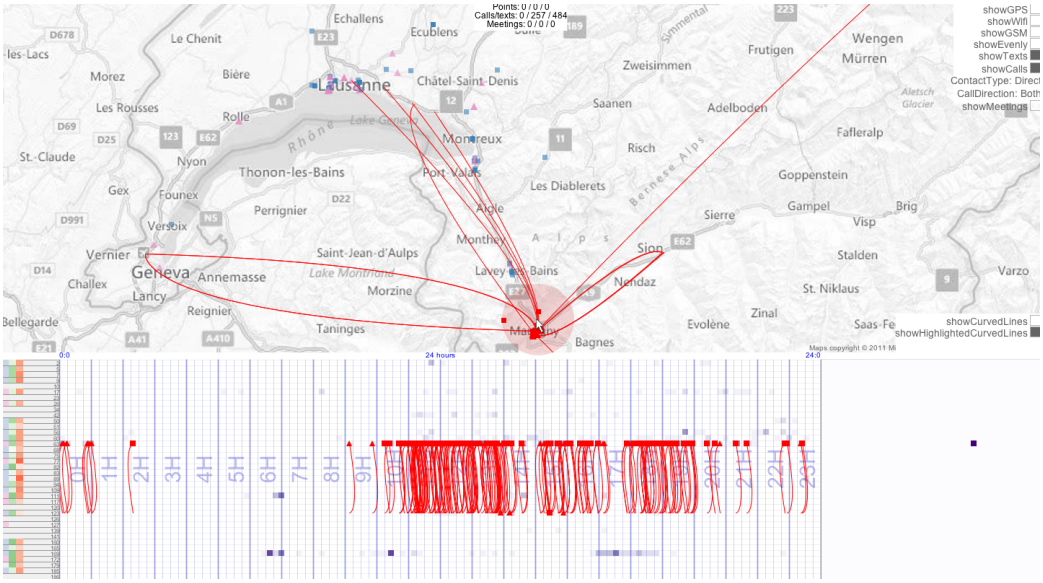

Figure 9: Most direct calls between participants involve one location and two participants (63 and 123), established by brushing the map with the mouse. The matrix is in 'density mode' and calls between these participants overwhelmingly dominate. Also note gaps in the reciprocal call record. This reveals internal inconsistencies in the data important for assessing their appropriate use.

In this case, it is consistent with a view that our set of users is a subset from a well-connected network. It illustrates the problem with studying social networks from an incoherent sample of participants, because of the resulting fragmented social network.

6.5. Reorderable matrix

We use re-orderable matrices [4] to show the connectivity between participants whose rows align with the timeline rows. We use three measures of social network activity: direct calls between participants (as a matrix), calls to/from mutual contacts of participants (as a matrix) and all calls made by each participant (as a barchart). These are summarised as number of calls (Fig. 11, left), number of contacts (Fig. 12) and average call length and can be filtered by call type. Colours can be interactively rescaled to match the value-range of interest.

Fig. 11 (top left) shows that direct calls/texts between participants are sparse and that most participants only have direct contact with one other participant. Following the evidence for the incoherent sample of a larger

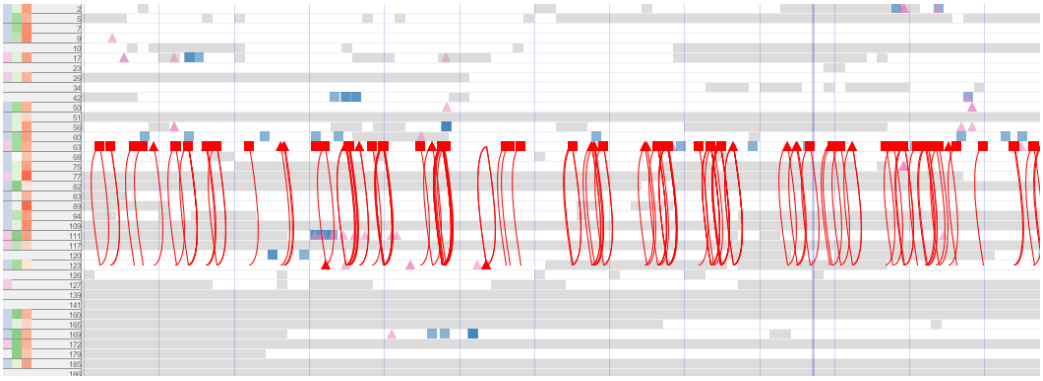

Figure 10: Zooming-in linear timeline for the same highlighted calls as in Fig. 9. Participant 123 (the lower participant) did not participate for much of the period, hence the reciprocal calls are not recorded.

social network in section 6.4, this matrix reveals that participants tend to call only *one* other contact in the study, suggesting that the sample we considered as incoherent has deliberately ensured that pairs of participants are included. Calls between mutual contacts (section 6.4) are much more numerous (Fig. 11, bottom right) and reveal a more nuanced set of social relations.

Sorting participants by their characteristics in the participant view, also sorts timeline rows and the columns and rows of the matrix, helping us identify broad usage patterns. Sorting by gender in Fig. 11 (top left) shows that of the direct calls to participants, most are between a female and male participant. The top middle matrix shows they are of a similar age. This is consistent with a view that the incoherent sample we have comprises couples in a relationship.

The matrix in Fig. 11 (top right) is sorted by social activity and suggests that more calls are made between more socially-active participants. This pattern is also broadly reflected in the bottom left matrix of calls to mutual friends. The bottom right matrix broadly shows that more socially-active participants share more mutual contacts.

The matrix view also serves as a means to highlight calls made between pairs of participants or mutual friends. A participant who calls himself is apparent in Fig. 11 (left; on the diagonal line; participant 60). Selecting this cell highlights the calls on the map and timeline. Full call details are available as a tooltip. None were georeferenced, but they occurred throughout the

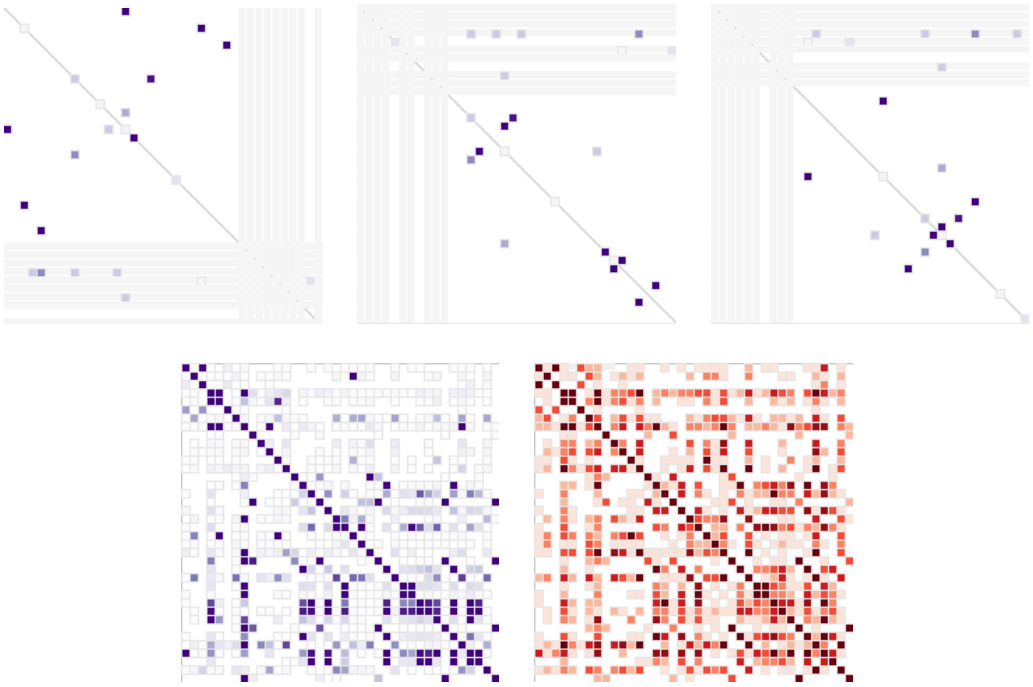

Figure 11: Social network matrices. *Top left:* Sorted by gender and coloured by number of calls/texts, showing that many calls are between people of the opposite gender. *Top middle:* Sorted by age and coloured by number of calls/texts, showing that people of a similar age group call each other. *Bottom left:* Sorted by social activity and coloured by the number of calls/texts made, showing that people with higher levels of social activity make more calls. *Bottom right:* sorted by social activity and coloured by calls to mutual contacts, showing that higher social activity is associated with calls to a wider diversity of people.

participation period as either incoming text messages or outgoing calls. We speculate that this participant may have been using voice and text messaging as reminders.

7. Other detectable forms of social contact

7.1. Colocation through time

By brushing on the timeline and map, we established that two participants – 15 and 56 – tend to visit the same places at the same time. Fig. 12 shows the timeline zoomed to the time they both travelled to Zurich together, arriving in the late evening and leaving in the early hours of the morning.

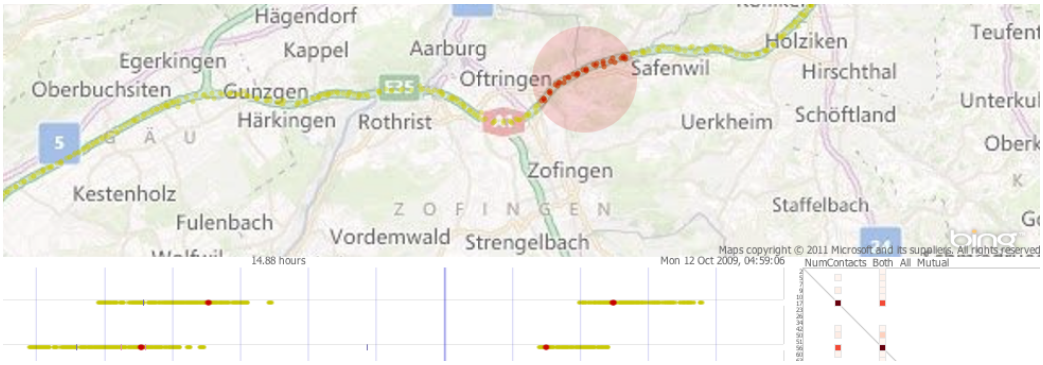

Figure 12: Two participants going to and from Zurich, an hour apart. We speculated that this unusual-looking behaviour could be due to participants being on public transport routes on services that are one hour apart, but on further investigation, we established that one of the participants had their device's timezone set incorrectly, highlighting the importance of checking data consistency.

Intriguingly, they appear to be travelling one hour apart on both outward and inward journeys. Investigating the GPS logs revealed that for the first few days of participant 17's logged data, the timezone was wrong by an hour, before being subsequently corrected. We speculate that this may have been after returning from a trip to a country in a different timezone.

7.2. Bluetooth

Each participant had a list of the bluetooth devices that their own device had been in proximity to. Since we knew the Bluetooth MAC addresses of most participants' devices, we were able to identify *when* participants were within close range and locate them using the method outlined in section 5.1. Fig. 13 shows such meetings between people, with one meeting highlighted. As expected, the matrix is less sparse than the calling matrix because proximity to someone is clearly more passive than calling someone.

As expected, when participants are collocated (as detected through Bluetooth), this happens during the day typically on weekdays; for some, this also includes the weekend. A few participants are collocated at all times. Further inspection confirms that the participants connected in Fig. 13 are of opposite genders and they appear to co-habit.

As with the other log data, there are gaps in the Bluetooth record and asymmetries where one device registers that it was 'seen' by another, but the reverse is is not necessarily true. This may be down to different Bluetooth

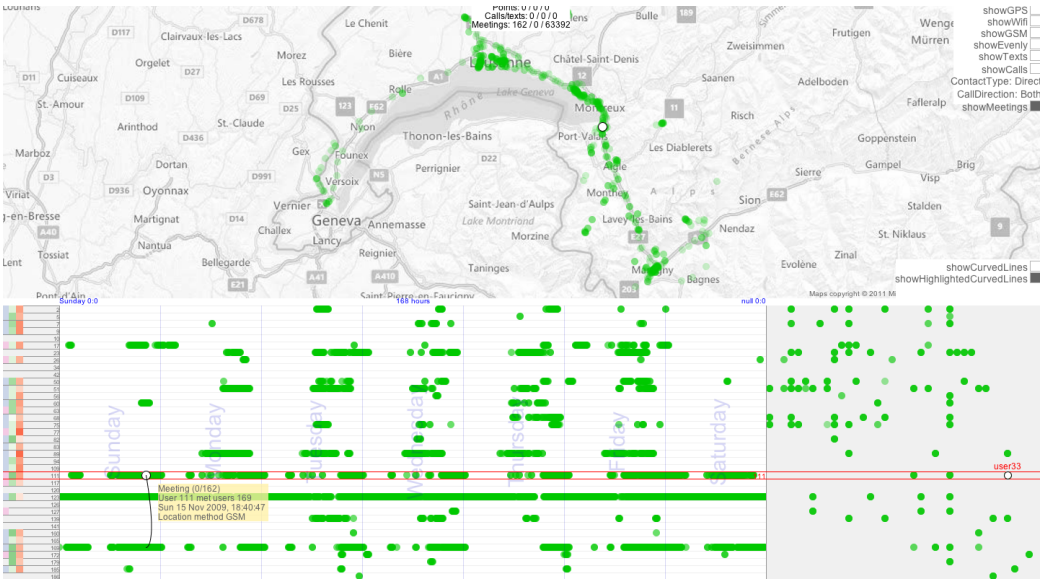

Figure 13: Close proximity between participants as detected by Bluetooth.

settings on devices, but is another example of inconsistencies we find in such data.

7.3. Findings about contact between participants

We have identified a few inconsistencies in the call log record, but by combining with temporal, spatial and other contextual data, we can determine who calls who and where they are when they do so. Other examples of collocated participants travelling together can be found. We can inspect individual calls and look at where and when they occur, and who they involve. There are numerous examples of couples who appear to cohabit, are of similar age and who call each other. More generally, using the re-orderable matrix, broad patterns in calling behaviour can be observed by age, gender and measures of social activity, although we would suggest that the sample is too small and too weakly connected to be able to make strong generalisations of this nature.

8. Research questions

We reflect upon our research questions and the extent to which we have been able to answer them.

8.1. To what extent can smartphone device logs help us to understand social communication behaviour?

The rich log data from smartphone devices enabled us to study participants and use context to help infer characteristics of their behaviour. It is useful to remember that social contact through email, online social media and face-to-face conversations is significant and is not recorded here. For this reason, we were particularly interested in spatial aspects of association, as this makes contact through smartphone devices less relevant; as such, associations derived through spatial proximity may be missed by simply studying call log records.

Various omissions and gaps discovered in these smartphone log files limited our ability to understand participants' social communication beyond inspecting specific cases. We expected that automatic logging on similar devices for participants who agreed to have their data shared, would produce a relatively complete record of behaviour. Although broadly the case, there were deviations from this – significant for our analysis – including a poor spatial record in some cases, short participation lengths, some with long gaps, a limited number of participants and sparse direct communication between participants (only 0.6% were direct). An implication of the gaps in the spatial record is that we were unable to georeference a third of calls and texts. Despite these deficiencies, in identifying and exploring calls and text messages between shared contacts external to the study - between 'mutual contacts' - we were able to identify a set of social communication behaviour, and learn about a wider social network, that was one degree away from the participants.

8.2. Can exploratory visualisation techniques help us characterise spatial and temporal aspects of participants' social networks?

We have found our exploratory visual analysis techniques to be powerful ways to characterise the data. Interactively-linked and brushable point- and density-based spatial, temporal and contact matrix views enabled us to gain an understanding of the phenomena the data represent and identify limitations to our analysis. Curved lines indicating calls allowed us to identify lack of reciprocal call records under some circumstances. Using the matrix to look at direct calls between participants and calls to and from mutual contacts outside the study, enabled us understand the structure of the participant sample we were working with. How well all this reflects spatial and temporal

characteristics of participants' social networks depends on how representative the data are. Our exploratory visual analysis techniques have helped us uncover data artefacts, some of which help us assess the suitability of the dataset for answering our questions, and some of which have given us insights into participants' social behaviour.

The techniques were designed for the scope of the data offered for the Data Challenge. As such, our methods may not scale well. In terms of speed and memory, we currently load 1.6 million GPS positions, 1.8 million wi-fi positions and 7.7 million GSM positions into memory and then regularly sample 11.7 million locations. More data may require a lower spatial sampling rate. In terms of visualisation scalability, having a row per individual will not scale up to many more participants. For more participants, other solutions would be needed.

8.3. Can linking spatial, temporal and call connectivity patterns help us explain how participants construct their social networks?

Our linking of spatial and temporal aspects of call connectivity has enabled us to study calling behaviour by place, time (by hour and day) and by individual. An example is the large number of calls made at one place (Fig. 9). Spatial and temporal signatures have helped us distinguish various types of location and activity, though Fig. 5 (bottom) suggests that some places contain both places of work and residential addresses. Importantly, studying the spatial nature of social networks enabled us to identify and describe social activity that could not have been detected through analysing participants' contact history alone. For example, the association between participants in Fig. 12 could not be derived from the call logs alone. Given that spatial proximity often reduces the need to communicate electronically, spatial context derived through explicit positioning technology or colocation using technologies such as Bluetooth, is an important aspect of association between participants that might otherwise be neglected.

Our approach was to show the data in as raw a form as possible; this means that the data we display most closely resemble the original data stored within the smartphone logs. We have demonstrated the substantial advantages to doing this; enabling data artefacts to be distinguished from behavioural artefacts. Nonetheless, some processing was required – to georeference the data calls and sample the spatial data at a regular temporal resolution, as described in section 5, and exploratory visualisation again enabled more informed decisions around how this might be done.

8.4. Can providing information about participants help us generalise our findings?

As mentioned, our ability to infer behaviours that might be generalised outside of this study was limited by the fact that the number of individuals participating in the study was small, direct social networks between participants were highly disconnected and some participants were present in the study for only a small amount of time. A more comprehensive sample along with a set of associated contextual information, would perhaps be needed for us to make firmer claims about identified behaviours. Even if a very large sample was achieved, however, our visual exploration of the data revealed many problems associated with these attribute-rich smartphone logs. If the same problems persist in many other smartphone datasets, then any research projects which attempt to make claims about behaviours might also be undermined. Since the object of many 'reality mining' studies is to develop algorithms for automatically inferring behaviours and characteristics from a small sample, with the implication that these algorithms could be scaled to a population, an appreciation of systematic errors in these data would be instructive.

We deliberately wanted to provide an interface to relatively raw data. This helped us understand limitations of the data and the next step might be to use more sophisticated modelling than our regular temporal sampling, to try to extrapolate more representative data.

8.5. Assessing and discussing data deficiencies

Although not one of our research questions, our fourth contribution (page 4) was to use visual exploration to assess data quality: how representative they are and their limitations for understanding the phenomena they represent. This is important because, as we have illustrated, there are numerous deficiencies with the data that may affect the validity of inferences made from them. These are present *despite* the protocols and agreements in place to ensure logs reflect participant activities as completely as possible [16]. For this reason, such deficiencies are likely to exist in similar datasets, particularly where participants' actions and locations can affect data quality, either intentionally (e.g. by turning off logging) or unintentionally (e.g. by not charging batteries or by being in an area with poor network availability).

The coordinated linking between space, time, participant and call log views of the data, coupled with interactive brushing, enabled us to establish data deficiencies that may affect the validity of inferences made from the

data. The call matrix indicated a sparsity of direct calls between participants and that most participants made and received direct calls to and from one participant. Using the matrix to show calls to individuals *outside* the study but with contacts *within* the study indicated that the sample of participants was probably part of a well-connected network of which we only had fleeting glimpses. These observations suggested that we only had a partial view of the social network and, as such, it would be difficult to make the wider observation about how participants use technology to construct social networks. Brushing calls on the timeline helped establish asymmetry in logged calls indicating that (for some reason) not all calls are logged. Since both incoming and outgoing calls should have been logged, we had not considered checking whether this was the case. Map and timeline brushing helped identify other inconsistencies in details, e.g. the one hour lag between two participants that eventually led to the discovery of the timezone configuration problem (Fig. 12). Although the latter example relied on serendipity to some extent, the flexible filtering and brushing amongst coordinated views helped uncover characteristics that one might not expect and therefore would not specifically look for. This demonstrates the value of flexible interactive visual exploration techniques.

9. Conclusion

Our flexible prototyping approach was able to incorporate and adapt interactive visualisation techniques for addressing our research questions relating to spatiotemporal aspects of social networks and our changing analytical needs. We were able to begin with map and timeline views, with interactive brushing of the map and timeline that were validated in a previous project, and then augment these with techniques that incorporate reorderable call matrices, proximity matrices and some contextual information about participants. These techniques and the flexibility with which we could adapt them, were successful in revealing the dataset's structure and helped us to address our research questions. They were also successful in allowing us to study characteristics of the data that impact interpretations regarding spatial aspects of social networks, including: widespread gaps in the logs, particularly in the spatial record; a poorly-connected sample of a wider social group characterised by few direct contacts most of which were between couples and many mutual contacts; and asymmetries in the reciprocal call records consistent with incomplete logging. We found that visual exploration – which has

the advantage of not relying on aggregation, clustering or machine learning – enabled us to qualitatively consider contextual information about participants. It also allowed us to understand the structure and limitations of our data. Even high-quality datasets regularly have deficiencies one might not expect, deficiencies that impact on the validity of interpretations that can be made from the data.

Whilst we were able to interactively explore different aspects of participants' behaviour and learn a great deal about specific participants, the small sample of 38 participants made it difficult to generalise behaviour. Despite this, we were able to find broad patterns in terms of gender, age and social activity, and the same visual techniques we use here have been shown to be highly effective at inferring more general structure from much larger datasets [3]. In addition, the high levels of communication activity that we found between shared participants external to the study – between 'mutual friends' – suggests that the sample of the Lausanne Data Collection Campaign is distinct and well-connected, and therefore that with a slightly larger sample, meaningful interactions could be more fully characterised.

This work demonstrates that by combining and facilitating comparison of diverse attribute-rich data, visual analytics is effective in helping provide insights, supporting interpretations and – perhaps more importantly in this context – helping assess the suitability and limitations of a dataset. Using detailed processed data from the Lausanne Data Collection Campaign, we have been able to identify, study and contextualise participants' contact behaviour. A consequence of this analysis is that we found problems associated with these detailed data, which may be common to other smartphone datasets. The data issues that we describe here may be used to inform automated and more systematic approaches to detecting errors in such data.

Acknowledgements

We thank Nokia and the organisers of the Nokia Data Challenge for the opportunity to take part and for providing us with data from the Lausanne Data Collection Campaign [16].

10. References

[1] Y. Altshuler, M. Fire, N. Aharony, Y. Elovici, and A. S. Pentland. How many makes a crowd? on the evolution of learning as a factor of commu-

nity coverage. In *Proceedings of the 5th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'12, pages 43–52, Berlin, Heidelberg, 2012. Springer-Verlag.

[2] G. Andrienko, N. Andrienko, U. Demsar, D. Dransch, J. Dykes, S. I. Fabrikant, M. Jern, M.-J. Kraak, H. Schumann, and C. Tominski. Space, time and visual analytics. *Int. J. Geogr. Inf. Sci.*, 24(10):1577–1600, Oct. 2010.

[3] R. Beecham and J. Wood. Exploring gendered cycling behaviours within a large-scale behavioural dataset. *Transportation Planning and Technology*, [in press].

[4] J. Bertin. *Semiology of Graphics*. ESRI Press, 2010.

[5] C. Brewer and M. Harrower. Colorbrewer: Color advice for maps. `http://colorbrewer2.org`.

[6] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[7] G. Chittaranjan, J. Blom, and D. Gatica-Perez. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *Proceedings of the 2011 15th Annual International Symposium on Wearable Computers*, ISWC '11, pages 29–36, Washington, DC, USA, 2011. IEEE Computer Society.

[8] T. M. Do and D. Gatica-Perez. Human interaction discovery in smartphone proximity networks. *Personal Ubiquitous Comput.*, 17(3):413–431, Mar. 2013.

[9] J. A. Dykes and D. M. Mountain. Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and applications. *Comput. Stat. Data Anal.*, 43(4):581–603, Aug. 2003.

[10] N. Eagle and A. (Sandy) Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, Mar. 2006.

[11] B. Fry. *Visualizing Data*. Cambridge University Press, 2007.

[12] B. Fry and C. Reas. Processing. `http://processing.org/`.

[13] F. Girardin, J. Blat, and C. Ratti. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing Magazine*, pages 36–43, 2008.

[14] T. Kapler and W. Wright. Geotime information visualization. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pages 25–32, Washington, DC, USA, 2004. IEEE Computer Society.

[15] T. Kim, J. Blom, and J. Stasko. Exploring Complex Mobile Life through Lightweight Visualizations. pages 37–40, Bergen, Norway, 2011. Eurographics Association.

[16] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In *Proceedings Mobile Data Challenge by Nokia Workshop, International Conference on Pervasive Computing*, 2012.

[17] J.-K. Min, J. Wiese, J. I. Hong, and J. Zimmerman. Mining smartphone data to classify life-facets of social relationships. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, pages 285–294, New York, NY, USA, 2013. ACM.

[18] Nokia. Mobile data challenge. `http://research.nokia.com/page/12000`.

[19] Z. Shen and K.-L. Ma. Mobivis: A visualization system for exploring mobile data. In *Visualization Symposium, 2008. PacificVIS '08. IEEE Pacific*, pages 175–182, March.

[20] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–343, Washington, DC, USA, 1996. IEEE Computer Society.

[21] A. Slingsby. Supporting the visual analysis of the behaviour of gulls. `http://bit.ly/HLZcFo`.

[22] A. Slingsby, R. Beecham, and J. Wood. Visual analysis of social networks in space and time. In *Proceedings of the Nokia Data Challenge Workshop, Pervasive 2012*, Newcastle, UK, 2012.

[23] A. Slingsby and J. Dykes. Experiences in involving analysts in visualisation design. In *Proceedings of BELIV2012*, Seattle, WA, USA, 2012.

[24] M. Theus. Interactive data visualization using Mondrian. *Journal of Statistical Software*, 7(11):1–9, 2003.

[25] J. Thomas and K. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.

[26] J. Wood, A. Slingsby, and J. Dykes. giCentreUtils. `http://gicentre.org/utils/`.

[27] J. Wood, A. Slingsby, and J. Dykes. Designing visual analytics systems for disease spread and evolution: Vast 2010 mini challenge 2 and 3 award: Good overall design and analysis. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 285 – 286. IEEE, 2010.

[28] J. Wood, A. Slingsby, and J. Dykes. Visualizing the dynamics of london's bicycle hire scheme. *Cartographica*, 46(4):239 – 251, 2011.