HAL open science

Combining sensors and surveys to study social contexts: Case of scientific conferences

Mathieu Génois, Maria Zens, Marcos Oliveira, Clemens M Lechner, Johann Schaible, Markus Strohmaier

To cite this version:

Mathieu Génois, Maria Zens, Marcos Oliveira, Clemens M Lechner, Johann Schaible, et al.. Combining sensors and surveys to study social contexts: Case of scientific conferences. Personality Science, 2023, 4, pp.e9957. 10.5964/ps.9957 . hal-03947154v2

HAL Id: hal-03947154

https://hal.science/hal-03947154v2

Submitted on 21 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

PERSONALITY Science

PsychOpen GOLD

Projects & Data

Combining Sensors and Surveys to Study Social Interactions: A Case of Four Science Conferences

Mathieu Génois[1], Maria Zens[2], Marcos Oliveira[3], Clemens M. Lechner[4], Johann Schaible[5], Markus Strohmaier[6]

[1] *Aix Marseille Univ, Université de Toulon, CNRS, CPT, Marseille, France.* **[2]** *GESIS - Leibniz-Institut für Sozialwissenschaften, Köln, Germany.* **[3]** *Computer Science Department, University of Exeter, Exeter, United Kingdom.* **[4]** *GESIS - Leibniz-Institut für Sozialwissenschaften, Mannheim, Germany.* **[5]** *Faculty of Computer Science and Engineering Science, TH-Köln, University of Applied Sciences, Gummersbach, Germany.* **[6]** *Business School, University of Mannheim, Mannheim, Germany.*

Personality Science, 2023, Vol. 4, Article e9957, https://doi.org/10.5964/ps.9957

Received: 2022-07-21 • **Accepted:** 2023-01-14 • **Published (VoR):** 2023-06-07

Handling Editor: Lauren Human, University of British Columbia Okanagan, Kelowna, Canada

Reviewing: *Round 1* - Anonymous #1; Anonymous #2. Open reviews are available. [see Index of Supplementary Materials]

Corresponding Author: Mathieu Génois, Centre de Physique Théorique, Campus de Luminy, Case 907, 163 Avenue de Luminy, 13288 Marseille Cedex 9, France. E-mail: mathieu.genois@cpt.univ-mrs.fr

Supplementary Materials: Materials [see Index of Supplementary Materials]

Abstract

We present a unique collection of four data sets to study social behaviour, collected during international scientific conferences. Interactions between participants were tracked using the SocioPatterns platform, which allows collecting face-to-face physical proximity events every 20 seconds. Through accompanying surveys, we gathered extensive information about the participants: sociodemographic characteristics, Big Five personality traits, DIAMONDS situation perceptions, measure of scientific attractiveness, motivations for attending the conferences, and perceptions of the crowd. Linking the sensor and survey data provides a rich window into social behaviour. At the individual level, the data sets allow personality scientists to investigate individual differences in social behaviour and pinpoint which individual characteristics (e.g., social roles, personality traits, situation perceptions) drive these individual differences. At the group level, the data allow to study the mechanisms responsible for interacting patterns within a scientific crowd during a social, networking and idea-sharing event.

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, CC BY 4.0, which permits unrestricted use, distribution, and reproduction, provided the original work is properly cited.

Keywords

face-to-face contacts, behavioral study, quantitative sociology, computational social science, social network, sociophysics, complex systems, complex networks

Relevance Statement

We present the first datasets that collect both quantitative measurements of human behaviour and information about the social and psychological characteristics of the participants, allowing for many explorations of the mechanisms of social systems.

Key Insights

- Longitudinal study of social interactions during scientific conferences
- Combines multiple data sources: sensor data and self-report surveys
- Correlation between behaviour and social/psychological dimensions of individuals
- Comparison between established, monodisciplinary and young, interdisciplinary scientific events

The study of human behaviour now includes device-based quantitative methods enabling researchers to track behaviour in unprecedented detail. Many of these novel methods have emerged with the expansion of electronic and online media, particularly mobile phones and the Internet. At the same time that such media increasingly shape our interpersonal behaviour, they provide us with the means to collect fine-grained data about human behaviour, which can be used for answering research questions.

The area in which these novel methods have arguably had the greatest impact is the study of social behaviour. Social behaviour comprises a broad class of behaviours, all of which involve some form of interaction and mutual influence among individuals. The APA Dictionary of Psychology (APA, 2022) defines human social behaviour as an "action that is influenced, directly or indirectly, by the actual, imagined, expected, or implied presence of others". In the present paper, we focus on a specific aspect of social behaviour, namely, on social interactions in situ as proxied by physical proximity. Whereas social interactions have traditionally been difficult to study outside of controlled laboratory settings, the ubiquity of mobile phones enables us to understand how individuals interact (Calabrese et al., 2015), and researchers have used it as a proxy for individuals' geographical location to investigate spatial crowd dynamics (Calabrese et al., 2015; Rojas et al., 2016). Likewise, the diffusion of GPS as an everyday tool was another step in the development of methods to probe human travel patterns (Rout et al., 2021; Sila-Nowicka et al., 2016). Remarkably, the Internet and its multiple usages have introduced new tools that open another window on human behaviour (e.g., online social networks, instant messaging, web browsing). In all these tools, the common feature is that they generate digital traces, data about the users' behaviour that can be automatically collected and stored.

Computational social scientists have rapidly noticed how they could use these data sources to study human behaviour, particularly data about online behaviour. Because of its vast availability, online data have been extensively used to investigate human behaviour. However, such research efforts have the caveat that results obtained with online media may not necessarily be transposed onto their real-world counterparts (Mellon & Prosser, 2017). Crucially, new questions arise; for example, do electronic and verbal communications share common properties? Are social circles similar online and offline? How do online and offline behaviour translate and impact one another?

To tackle these and other critical questions, we need to be able to probe the real world in the same quantitative way as the online world. To that end, researchers have developed sensors, either relying on existing infrastructure—usually smartphones (Stopczynski et al., 2014; Vu et al., 2010)—or designing their own (Choudhury & Pentland, 2003; Salathé et al., 2010). Such sensors detect physical proximity between participants, which constitutes a proxy for social contacts (Malik, 2018; Schaible et al., 2022). This redefinition of behaviour measurement allows for collecting quantitative information about how individuals interact with each other in physical space.

We have thus a new tool to study human social behaviour in situ, which gives new ways to look at the phenomena at play. These interactions are proxied by individuals' physical proximity in space, as registered by the sensors, and can be analysed on both the individual and group level. Such interactions in situ capture the more objective, physical side of social behaviour, which is however always complemented and enriched in our data collections by additional information on social roles, traits, and motives measured by surveys. By pairing sensor data with surveys, we can objectively measure, and quantify individual differences in, social behavior. Moreover, we can pinpoint individual and contextual characteristics that shape social behaviour and underlie individual differences therein. We can thus contribute to a better understanding of the linkage between personality and social behaviour, a topic that has received considerable attention in personality science in recent years (e.g., Back, 2021; Breil et al., 2019).

In this work, we focus on scientific conferences as an example of a social context where social interactions can be driven by several factors: social roles and social status, personality traits, situation perceptions, and motivations to cite but a few. We chose scientific conferences for the relative simplicity of the situation, with individuals confined to a well-defined space, free to interact and with synchronised schedules of high (breaks) and low (talk sessions) activity periods. Additionally, conferences as a social event are of substantive interest in their own right to several fields of research, such as applied psychology, sociology of science, social psychology. Finally, they had the added benefit of maximising participation rates as scientist are more likely to agree to take part in experimental studies, and being convenient to monitor as our team was the main organiser of the events. Such data sets allow for a wide range of exploratory studies regarding the effect of each of these factors on contact behaviour, correlations between

them, insights into crowd dynamics in the sociology of science, and general properties of contacts between individuals in different contexts.

For example, the sociodemographic attributes are interesting in the perspective of a sociology of science: one can investigate how researchers with different academic statuses interact in the context of a conference; or, taking advantage of the interdisciplinarity of three of the studied cases, how researchers with different backgrounds mix at a common event. The data can provide insights in, e.g. disciplinary openness and cohesion, the role of academic hierarchies, and it can help to detect biases in communication patterns and group formation. Using the results of the personality, situation and motivation questions, one can look into the relations between these components and behaviour as measured by the sensors. Finally, the perception gap experiments, in which participants were asked to estimate the size of selected sociodemographic groups in the crowd, allow for a comparison between the evaluation of a social situation by individuals and the reality of their interactions.

The aim of the paper is to present the data collected during a set four conferences, not the data collection method, which is presented for transparency. General information about the method is available in Schaible et al. (2022) and Kontro and Génois (2020). Researchers interested in the SocioPatterns equipment can refer to the collaboration website (www.sociopatterns.org); it must however be noted that the equipment belongs to the SocioPatterns collaboration and is not freely available.

Method

This section presents all details of the data collection procedure.

General Description of the Events

The data were collected during four events organised by GESIS, the Leibniz Institute for Social Sciences, in Cologne, Germany. Throughout the manuscript, we refer to them using the following labels:

- **WS16:** The 3rd GESIS Computational Social Science Winter Symposium, held on November 30 and December 1, 2016. This event was part of a series on computational social science organised by GESIS. This edition had the specific topic of "Understanding social systems via computational approaches and new kinds of data".[1]

1) GESIS CSS WinSymp (2016). 3rd GESIS Computational Social Science Winter Symposium: Understanding social systems via computational approaches and new kinds of data, Cologne, Germany. http://www.gesis.org/css-wintersymposium/home/ (https://web.archive.org/web/20160923120609/), accessed: 2022-05-26.

- **ICCSS17:** The International Conference on Computational Social Science, held from July 10 to 13, 2017. Broadly speaking, the conference is known for bringing interdisciplinary researchers together for advancing social science knowledge through computational methods.[2]
- **ECSS18:** The Eurosymposium on Computational Social Science, held from December 5 to 7, 2018. This event was part of the European Symposium Series on Societal Challenges in Computational Social Science. This edition had the headline of "Bias and Discrimination".[3]
- **ECIR19:** The 41st European Conference on Information Retrieval, held from April 14 to 18, 2019. The conference is the European forum for the presentation of research in the field of Information Retrieval.[4]

Though all events occurred in Cologne, Germany, they were organised in different locations. WS16 was held at the KOMED convention centre at MediaPark[5], whereas ICCSS17, ECSS18, and ECIR19 took place at the Maternushaus hotel[6].

The first three conferences (i.e., WS16, ICCSS17, and ECSS18) were interdisciplinary, gathering researchers from Social Sciences, Computer Sciences and Natural Sciences. In contrast, ECIR19 was focused on the Computer Science field. For the last three conferences (i.e., ICCSS17, ECSS18, and ECIR19), the first day consisted of a separate workshop/pre-symposium day, for which contact data was also gathered except for ECSS18, for which we have contact data only for the main conference on December 6 and 7. Full contextual information about the conferences and the venues, which may be useful to researchers wishing to examine how context (*e.g.*, timing of breaks) relates to social behaviour, is available online (see the footnotes for access to the original and archived versions of the conference websites and venue websites).

Table 1 lists basic statistics of participation to the studies. Overall, we have a very high participation rate, with more than 70% of attendees partaking in the studies. In the case of contact data, we have excellent coverage of the conferences' crowds; it is greater than 90% for three studies and 80% for ECSS18. The survey response rate is also good. We have at least partial information for more than 70% of the studied population.

2) IC2S2 (2017). The 3rd Annual International Conference on Computational Social Science, Cologne, Germany. https://ic2s2.org/2017/ (https://web.archive.org/web/20170617162339/), accessed: 2022-05-26.

3) GESIS CSS EuroSymp (2018). European Symposium Series on Societal Challenges in Computational Social Science: Bias and Discrimination, Cologne, Germany. http://symposium.computationalsocialscience.eu/2018 (https://web.archive.org/web/20181118085933/), accessed: 2022-05-26.

4) ECIR (2019). 41st European Conference on Information Retrieval, Cologne, Germany, http://ecir2019.org:80/ (https://web.archive.org/web/20190305021221/), accessed: 2022-05-26.

5) KOMED Zentrum für Veranstaltungen, https://www.komed-veranstaltungen.de/en/, accessed: 2022-05-26.

6) Maternushaus, https://tagen.erzbistum-koeln.de/maternushaus/start/, accessed: 2022-05-26.

Table 1

Statistics of Participation to the Studies

Study	WS16	ICCSS17	ECSS18	ECIR19
N	149	339	211	270
N_p	144 (96.6%)	284 (83.8%)	205 (97.2%)	190 (70.3%)
N_p^*	144 (96.6%)	277 (81.7%)	171 (81.0%)	178 (65.9%)
N_c	138 (95.8%)	274 (96.5%)	164 (80.0%)	172 (90.5%)
N_d	122 (83.3%)	213 (75.0%)	155 (75.6%)	140 (73.7%)

Note. N is the total number of participants to the conference; N_p is the number who agreed to take part in the study; N_p^* is the number for which we have data (contact and/or survey); N_c is the number for which we have contact data; N_d is the number for which we have at least partial sociodemographic information. Percentages for N_p and N_p^* are calculated with respect to N; percentages for N_c and N_d are calculated for the studied population and thus with respect to N_p.

Contact Data

The SocioPatterns Platform

The first part of each study consists in recording interactions between participants. A *social interaction* can include many different behaviours, such as conversation, physical contact, and eye contact. All are relevant for the analysis of ties within a crowd. In the present case, we focus on the more straightforward, broader definition of a *contact* as a physical, face-to-face proximity event. Although physical proximity between individuals does not necessarily imply an interaction, previous work shows that this signal constitutes an excellent proxy, which enables the analysis of the structure of a social context (Schaible et al., 2022).

We used the SocioPatterns platform (Cattuto et al., 2010) to collect contacts between participants, which has been largely used in the past decade to explore interaction patterns in social contexts (Génois & Barrat, 2018; Kiti et al., 2016; Kontro & Génois, 2020; Oliveira et al., 2022; Ozella et al., 2021; Vanhems et al., 2013). This equipment consists of sensors attached to the participants' name tags and antennas covering the conference venue to collect contact data from the sensors. Each sensor carries an RFID chip and can detect other sensors in the vicinity within a ~1.5 m radius. Furthermore, as the human body blocks the emitted signal, detection only occurs when two individuals are face-to-face (i.e., in their respective front half-spheres). An event with such proximity and geometry defines a *contact*. Contacts are recorded every 20 seconds and are limited to 40 simultaneous contacts for each individual within a 20-seconds time window. By design, contacts lasting at least 20 seconds have ~100% chance of being recorded. Shorter contacts may be recorded, with a probability decreasing as their duration decreases.

Contact detection does not depend on the orientation of the sensor: As the name tag does not block the signal emitted by the sensors, the detection occurs whether the

name tag is backwards or not; this ensures that data is collected even if the name tag is backwards. However, it may happen that the name tag does not stay on the chest of the person, for instance when participants have it on their back or keep it attached to a pocket, belt, etc. Furthermore, some participants may forget their sensor from one day to the next, or remove it for a time and leave it unattended. Ultimately, all those events lead to some wrong detection of contacts, which generates noise in the data. Controlling for such events is impossible in the setting. However, this limitation does not make the data invalid, as shown in Elmer et al. (2019). Furthermore, the network science literature of the past decade shows unequivocally that relevant information about social structures can be extracted from such data (see for example Stehlé et al., 2013 about gender homophily in a primary school, or Mastrandrea et al., 2015 for a comparison between sensor data, surveys and online ties).

Setting up the Contact Tracking Platform

As sensors only have limited memory, antennas are necessary to collect the data from them continuously. Coverage of the conference venue is thus crucial to ensure that the maximum amount of contacts is collected. Antennas have a theoretical detection radius of ~30 m. Thus, we examined each conference venue floor plan to identify the suitable number of antennas needed. Because sensors and antenna communicate via radio waves, we performed tests *in situ* to evaluate the impact of obstacles, in particular walls and windows which may block the signal. Antennas were thus positioned in order to minimise the data loss. See the Supplementary Materials for a detailed description of the coverage of each venue.

By design, contact detection occurs only on the area covered by the antennas. Thus, no contact detection can occur outside the conference venue. The data therefore does not include interactions that happened during social events or informal meetings that took place outside.

Broadly speaking, RFID sensors are inexpensive but deploying them requires some specialised knowledge & experience. The most limiting factor is time and manpower: setting up the data collections presented here has necessitated a team of 4 to 6 persons each time, including at least one expert in SocioPatterns studies to ensure the proper functioning of the platform and the validity of the setup. Setting up the survey required a server to allow for online answering.

The equipment for the data collections belongs to the SocioPatterns collaboration and its sharing is limited. Similar studies have been done through other types of sensors (for example using Bluetooth from smartphones) which price, usability and versatility vary (see Schaible et al., 2022). Should researchers be interested in such a study, the authors are available for discussion.

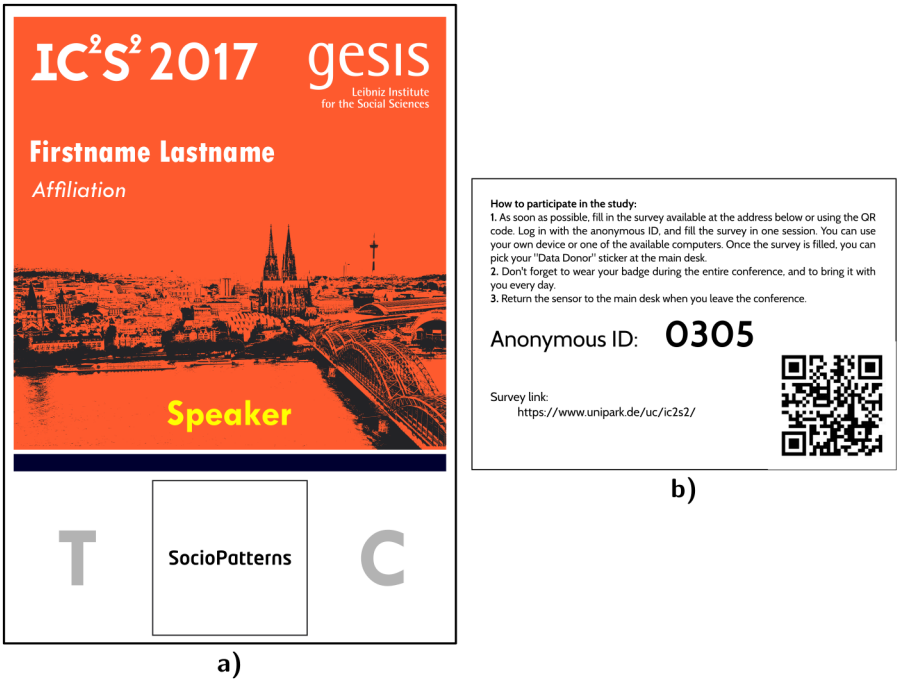
Participation and Sensor Distribution

Participation was offered to all attendees of the conferences upon registration (usually online before the event); attendees could opt out at their arrival at the event. Table 1 summarises the resulting participation rates.

To avoid manipulation by the participants, we preemptively installed sensors within the name tags used for the conferences (see Figure 1a). Before the conference, we sent an e-mail to all participants informing them that a SocioPatterns study was taking place during the conference, attached with a form of consent with a complete description of the data collection (see Supplementary Materials).

Figure 1

Example of Name Tag (a) and Survey Filling Information (b) Provided to the Participants



Note. a) The square frame at the bottom was reserved to receive the SocioPatterns sensor. The name tag is 105x148 mm, the sensor frame is 36x36 mm. b) When handed over to participants, each name tag contains an envelope with the information to fill out the survey, the anonymous ID used to connect the survey data to the contact data, and the link to the survey both written and as a QR code.

No compensation was offered for the participation. Upon registration at the conference, participants could choose to participate or refuse. A data collection team member was

also available to answer questions. If they agreed, they were given a form of consent to sign. If they refused, the sensor was removed from the name tag. When leaving at the end of a conference day, participants kept their name tags with the sensor and brought them back the next day. We note that no contact detection occurs outside the conference venue. Upon leaving the conference permanently, the participant returned their name tag to the registration desk.

Data Cleaning

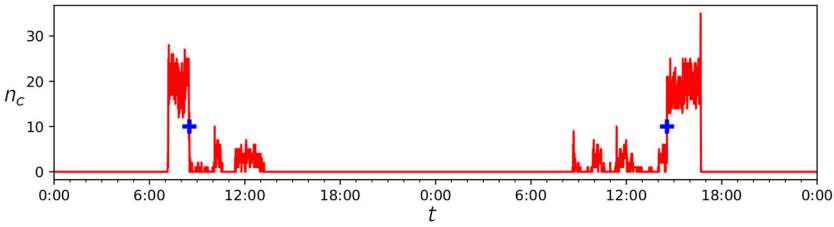
The raw data gathered by the antennas first went through a preprocessing phase, in which the contacts are *aligned*. This process was necessary because neither sensors nor antennas include an internal clock. Thus their data had to be synchronised. Furthermore, the data were *binned* into 20 seconds time windows.

In all four conferences, we used the same setup to be able to detect the precise moments when the sensor was handed over to the participant and returned to us (similar setup as in Kontro & Génois, 2020). As sensors are functioning continuously when they are powered, when all sensors are stored together they constantly detect each other, which results for each sensor in a very high number of simultaneous contacts. This level of activity is blatantly different from the situation where the sensor is deployed, during which the number of simultaneous contacts is relatively low (usually under 10). The sharp difference in activity level between these two situations allows us to very easily detect the moment a sensor is removed from storage and given to a participant, hence to determine the distribution time for each sensor. Similarly, when a sensor is returned to the storage the activity level jumps, which can be as easily detected and gives the return time of the sensor (see Figure 2a).

In practical terms, we did not distribute a set of name tags and listed their identifiers as beacons. We left a sufficiently large number of beacons in the returning box, allowing us to detect distribution and return times based on the jumps in the number of contacts detected by each sensor. For each sensor, we deleted all contacts recorded before distribution and after the return. Finally, all sensors that were not used in contact detection—beacons and undistributed name tags—were removed from the data.

Figure 2

Example of Distribution and Return Time Detection for Cleaning the Contact Data (a) and Resulting tij File (b)



a)

```
1480486100 89 79
1480486100 35 79
1480486100 56 91
1480486120 56 79
1480486160 89 79
1480486180 56 18
1480486200 56 18
1480486200 35 79
1480486220 56 18
1480486240 56 18
```

b)

Note. a) The number of contacts n_c (red line) detected by a sensor before distribution and after return are significantly higher than during the study. By simply using a threshold at n_c = 10 contacts per time step, we are able to precisely detect the time this sensor was distributed and returned (blue crosses). b) This example from the WS16 dataset lists the first 10 contacts recorded, occurring between time 1480486100 and 1480486240. The first line indicates that the contact occurred between participants 89 and 79 at time 1480486100.

Data Formatting

After preprocessing and cleaning, the resulting data is a *temporal network* in which the nodes are the participants, and the links represent contacts, appearing and disappearing as time passes. The contact data was formatted as *tij* file (see Figure 2b). Each line of the file corresponds to one contact occurring at time t between nodes i and j. Time stamp t is given as a standard UNIX Epoch time (*i.e.* number of seconds since January 1st, 1970). Contacts are ordered according to time; all contacts occurring simultaneously are thus gathered at the same place in the file.

Because of the time binning, all time stamps are multiples of 20 seconds, and each reported contact is considered to have lasted 20 seconds. Continuous interactions (*i.e.* contacts that occur between the same two participants on several consecutive time bins) are not reported as such and must be reconstructed from the 20 seconds contacts that constitute them.

For example, in Figure 2b the first line indicates that a contact occurred between participants 89 and 79 at time 1480486100, which corresponds to November 30, 2016 at 07:08:20. Lines 6, 7, 9 and 10 indicate that four consecutive contacts occurred between nodes 56 and 18, constituting an interaction which started on time 1480486180 (November 30, 2016 at 07:09:40) and lasted 80 seconds.

Surveys

Organisation & Data Anonymity

In addition to the contact data, we used surveys to gather information about the participants. These self-administered online surveys were available at the beginning of the first day of the conferences. Participants were asked to complete them as soon as possible and typically completed them upon arrival at the venue or within a few hours after their arrival. To distinguish participants who completed the survey only partially from participants who did not take the survey, in the survey data missing answers were labelled "NA" in the first case (partial completion) and left blank in the second case (no survey data).

To link the contact data with the survey data while ensuring anonymity, we used a system of anonymous identifiers (IDs). Each sensor has its ID consisting of four numerical digits, which uniquely identify it in the contact data. Along with the name tag, each participant was given an envelope containing this identifier to be used as their identifier when answering the survey (see Figure 1b). Because this anonymous identifier (in the envelopes and sensors) does not have any personal information, we ensure the anonymity of the participants. The anonymous IDs were further replaced by random numbers in the final data, ensuring that no link between the data collection and the final data could be established.

Content

The surveys consisted of several sections, covering different axes of inquiry that are relevant to personality science: respondents' sociodemographic characteristics (broadly defined and also including, for example, their disciplinary background and roles at the conference), personality traits (Big Five model; John et al., 2008), situation perceptions (DIAMONDS model; Rauthmann et al., 2014), scientific attractiveness, motivations to attend the event and perception gap regarding the gender distribution of the crowd. Table 2 summarises the content of the survey for each conference.

Table 2

Axes of Study for Surveys

Axis	WS16	ICCSS17	ECSS18	ECIR19
Sociodemographic characteristics	x	x	x	x
Age group	x	x	x	x
Gender	x	x	x	x
Age of the oldest child				x
Country of residence	x	x		
Primary language	x	x	x	x
Academic status	x	x	x	x
Disciplinary background	x	x	x	x
Role in the conference	x	x	x	x
Participation to a previous conference	x	x	x	x
Participation to the pre-symposium			x	x
Lunch choice				x
Number of persons known at the conference		x	x	x
Personality	x	x	x	x
Big Five personality traits	x	x	x	x
Personality facets	x	x		
Situation perception (DIAMONDS)			x	x
Scientific attractiveness		x	x	x
Self rated attractiveness		x	x	x
Number of citations (personal)		x	x	x
Number of citations (other participants)				x
Number of citations (closest peers)				x
Motivations to attend			x	x
Perception gap			x	x
Share of female participants			x	x
Share of professors			x	x
Share of participants younger than 30				x
Share of German-speaking participants				x

In all four conferences, we investigated participants' sociodemographic characteristics; however, the list of items was not always the same. We dropped the question about the country of residence after finding it not relevant. After WS16, we added questions about the number of persons in the conference that participants knew before the event and the number of citations, in parallel with scientific attractiveness, to investigate potential mechanisms for connecting behaviour. In the case of ECSS18 and ECIR19, these events had a pre-symposium, so we asked about participation in these activities. Finally, for ECIR only, we added questions about lunch options (for organisation purposes) and the

number of citations of other participants and peers to have insight into how participants see themselves concerning the crowd and their peers.

The second part of the study concerns personality traits, which we assessed using the established Big Five model (John et al., 2008). In the first two conferences, we administered the 30-Item BFI-2-S (Soto & John, 2017), which allows investigating 15 narrow personality facets in addition to the Big Five domains (Openness, Conscientiousness, Extraversion, Agreeableness, and Negative Emotionality). In later conferences, we opted for shorter Big Five instruments, namely the 15-item BFI-2-XS (Soto & John, 2017) for ECSS18 and the 10-item BFI-10 (Rammstedt & John, 2007) for ECIR19, to make space for other items in the survey. The ultra-short BFI-2-XS and BFI-10 allow for an exploration of Big Five domains but not facets.

To broaden the space of individual-differences constructs assessed, at ECSS18 and ECIR19 we added a measure of situation perceptions as conceived in the DIAMONDS model (Duty, Intellect, Adversity, Mating, pOsitivity, Negativity, Deception, Sociality; Rauthmann et al., 2014). Situation perceptions refer to how people perceive and construe situations, including the situation's action imperatives. To measure these situation perceptions, we slightly adapted the S8-III, an ultra-short scale measuring each of the eight dimensions with one item (Rauthmann & Sherman, 2016). We reworded the introduction such that it referred to the specific situation of scientific conferences and slightly changed some items to align them with the context and target population being studied.

With the scientific attractiveness axis, we aim to understand whether respondents' scientific status is relevant to understanding contact behaviour. Depending on the conference, we assessed scientific attractiveness in terms of perceived status but also several factual measures such as number of citations.

The motivations axis contains a simple question about the participant's motivations to attend the conference. This axis complements personality traits and the more generic DIAMONDS situation perceptions: it aims at understanding whether behaviour in such contexts is more directed by the nature of the participants or by their intentions.

Finally, the perception gap axis gathers questions about how the structure of the crowd is perceived by the participants in terms of the size of minorities/majorities. This information can inform us about disparities in perception, which can then be correlated with the social network structure as given by the contact data.

For a detailed description of the questions for each survey, the codebooks and questionnaires of the surveys are available with the contact data (see following section).

Transparency, Openness, and Reproducibility

Pre-registration

The studies are exploratory and thus were not pre-registered.

Hypothesis Testing

The aim of the present paper is only to present the collected data and does not test any hypothesis.

Data

The contact data are available in GESIS's SowiDataNet|datorium at the following link: https://doi.org/10.7802/2351

For privacy reasons, the raw contact data (i.e. data before preprocessing and cleaning as gathered by the antennas) are not available, as it contains the sensor IDs that were used during the data collection. For privacy reasons and to comply with the legal regulations concerning the collecting, use and sharing of personal data (GDPR), the complete survey data are available only through direct request to Mathieu Génois (mathieu.genois@cpt.univ-mrs.fr). The sharing of these data requires the signature of a sharing agreement that imposes several restrictions, in order to prevent inappropriate uses of the data. Excerpts of the survey data are however available along with codebooks, questionnaires and forms of consent at the following link: https://doi.org/10.7802/2352

This excerpt contains the information about Age class and Gender for WS16 and ICCSS17, Age class only for ECSS18 and ECIR19.

In order to comply with legal regulations about data use, access to both the contact and the survey data is restricted to scientific purposes only.

Scripts, Code, Syntax

The code for the extracting and preprocessing of the raw data gathered by the antennas is not available, for proprietary reasons. The program to produce Table 3 and Figures 3 and 4 is available in the Supplementary Materials. It relies on the tempnet library available at: https://github.com/mgenois/RandTempNet

Table 3

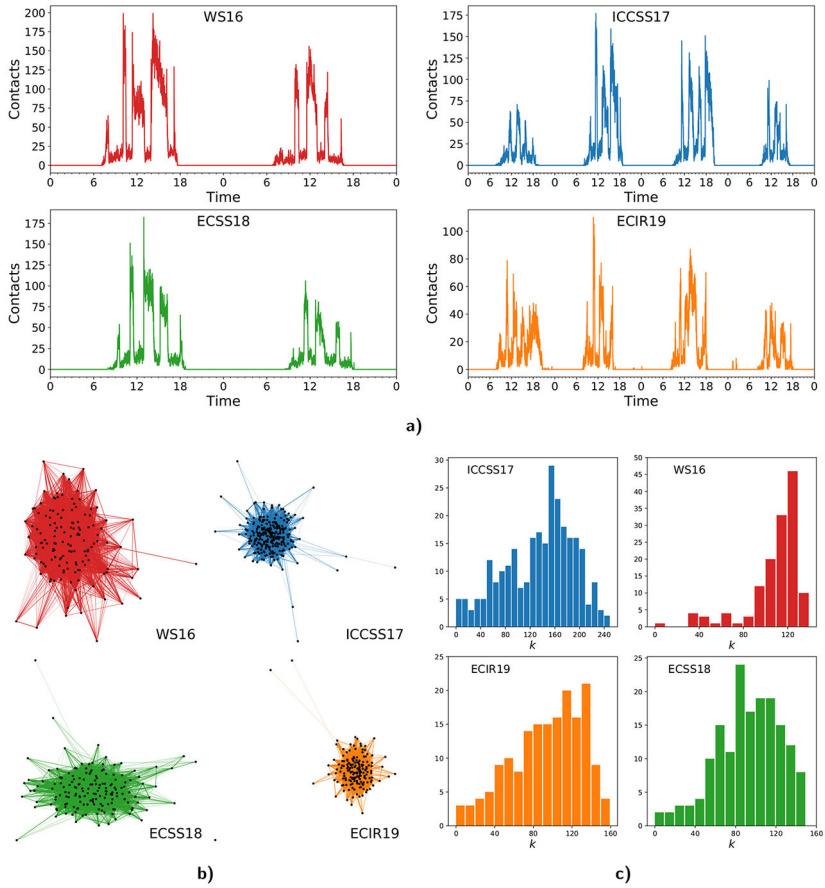
General Properties of the Contact Networks

Study	WS16	ICCSS17	ECSS18	ECIR19
C	153 371	229 536	96 362	132 949
ρ	0.793	0.495	0.567	0.550
$<k>$	108.6	135.2	92.4	94.1
$<c>$	0.868	0.694	0.717	0.746

Note. C is the total number of instantaneous contacts recorded; ρ is the density of the aggregated network, *i.e.* the fraction of possible connections that occurred during the event; $<k>$ is the average degree of the aggregated network, *i.e.* the average number of persons one participant met during the event; $<c>$ is the average clustering of the aggregated network.

Figure 3

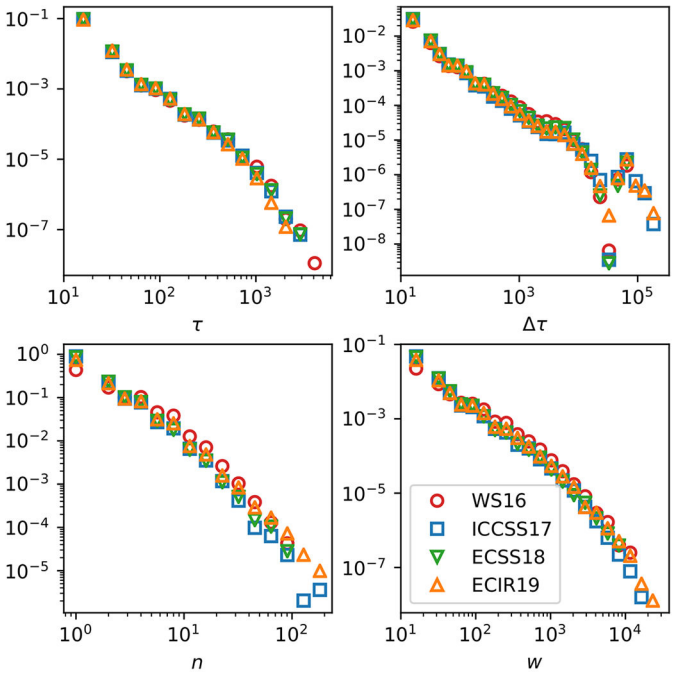
Typical Characteristics of the Contact Networks: Activity Timelines (a), Visualisation of the Aggregated Contact Networks (b), Degree Distributions (c)



Note. a) We plot the total number of contacts occurring in each 20 seconds time step. Curves exhibit the circadian rhythm (activity during the day and no activity at night) and alternating periods of social times (coffee breaks, lunch, poster session) and low activity windows (talk sessions). b) Nodes are individuals; a link exists between two individuals if they have been at least once in contact during the event; the width of the link is proportional to the total contact duration between the two individuals. Node positions were set using a spring layout, were links are equated to springs with a stiffness proportional to its weight; all networks exhibit the same absence of visible global structure. c) The degree k of an individual is the number of connection it has in the aggregated network, *i.e.* the number of individuals with whom it has interacted at least once during the event. All distributions are skewed towards high values, indicating a large mixing: most people have interacted at least once with a significant fraction of the crowd.

Figure 4

Distributions of Temporal Properties of Network Links



Note. The temporal properties of links are contact durations τ (continuous interactions between two participants), inter-contact durations $\Delta\tau$ (between two consecutive contacts), number of contacts n between two participants, total contact duration w between two participants.

Other Supplements

A Supplementary presenting the plans of the venues and where the data collections were performed and and example of a form of consent is available in the Supplementary Materials.

Results

This section presents general statistics of the data sets.

Properties of the Contact Networks

The temporal networks obtained through the SocioPatterns studies consist of temporal links that indicate, every 20 seconds, which participants are in contact. We denote C as the total number of these instantaneous contacts, which describes the overall recorded

activity in an event (see Table 3). This activity changes over time, so we further define *contact activity* as the number of instantaneous contacts occurring per time step. It describes the evolution of the interaction level between participants (Figure 3a). This evolution is similar for all conferences: we observe a circadian rhythm, with active days and inactive nights. Two phenomena are responsible for this property of the contact activity. First, collecting data only in the conference venue automatically limits the detection of activity to the period when participants are in. However, the general flux of participants in the venue in the morning and out of the venue in the evening is precisely the pattern (though rather trivial) our setup detects. The active periods thus exhibit a wave shape with a progressive increase at the beginning and a decrease at the end, modulated by the succession of high and low activity periods. High activity periods are "social times" such as registration, coffee/lunch breaks, or poster sessions; low activity periods are talk sessions.

To assess the dynamics of face-to-face interactions, we evaluated some basic statistics regarding the contacts (see Figure 4). We define any series of instantaneous contacts occurring sequentially without in-between gaps as one continuous contact with a duration of τ (*i.e.*, an interaction). With this definition, we can then explore the overall temporal properties of the interactions (*i.e.*, the distributions of τ). Additionally, we examined the inter-contact durations, denoted $\Delta\tau$, between two consecutive interactions between the same participants. Furthermore, we evaluated the number of contacts n and the total contact duration (*i.e.*, weight) w occurring between two participants. By examining the empirical distributions of these quantities, we found well-known, large-tail shaped distributions. This finding indicates that the most numerous contacts last 20 seconds, the most numerous inter-contact durations last 20 seconds, most pairs of participants interacted only once, and for one contact of 20 seconds only. However, extremely long instances of each of these properties also occur, with a small but not negligible probability, as indicated by the roughly power-law aspect of the distributions. Finally, the distribution of $\Delta\tau$ exhibits the usual depletion/inflation feature caused by the circadian rhythm in the activity data.

By flattening the temporal network across the temporal dimension, we obtained an *aggregated network* in which nodes are the participants, and a link exists between two nodes if the participants have interacted at least once during the event. We performed a standard analysis of these networks and found that all four are very dense (see Table 3). This finding is primarily because the venues were somewhat crowded, ensuring that each participant came into contact with a significant fraction of the rest of the crowd. One can indeed see from the visualisations of the networks that connections are very numerous (Figure 3b).

The degree of a node in the aggregated network indicates the number of participants it has been at least once in contact. The high density of the networks appears as well on

the degree distributions, which are skewed towards high values, indicating that, indeed, most participants interact at least once with most of the other participants (Figure 3c).

One key aspect of the contacts is the high density of the contact network (values close to 1 indicate that each participant had at least one contact with almost all the others) due to that fact that most interactions have a very short duration (20 seconds). This leads to a crucial question: how to distinguish socially relevant interactions from random physical proximities? This question remains currently unanswered; what can be said is that applying a threshold on the contact duration, though tempting, is not the way to go. Although the probability for a contact to be relevant increases with duration, Figure 4 shows that there exists no "natural" threshold in the distribution of contact duration that would indicate a change of nature in the interaction. Indeed, socially relevant interactions may be last a few seconds while irrelevant contacts may last more than a minute. Applying a threshold to reduce the probability of incorporating "irrelevant" interactions is a possibility, but any subsequent analysis must then include a robustness check which verifies that any observed phenomenon is robust to a change in the threshold value. As a consequence, in the present data we have chosen to give access to all interactions recorded, and leave the choice of a filtering method to researchers who will use the data.

Survey Information

The accompanying surveys assessing the axes shown in Table 1 were conducted as online surveys but administered on-site. After arriving at the conferences, participants were invited to participate in the survey, which they could fill out on laptops provided by the conference organisers or on their own devices. For linkage of the survey data to the sensor data, the first item of each survey always required participants to provide their sensor ID.

At WS16 and ICSS17, there was only one survey. Because some of the survey items might be reactive (*i.e.*, respond to the experiences at the conference), efforts were made to encourage participants to fill in the survey immediately after arriving at the conference—the majority of participants filled in the survey on the first conference day. At ECSS18 and ECIR19, participants were invited to participate in a second survey toward the end of the conference, in which additional questions that depended on participants' experiences during the conferences (especially about perception gap) were asked.

The survey participation rates relative to the number of participants who wore a sensor ranged from 73.7% in ECIR19 to 83.3% in WS16, as shown in Table 2. Item missingness among those who started the survey was negligible (typically < 5%) at all conferences. The length of the surveys was kept short to avoid interfering with other conference activities and to minimise respondent burden. Respondents typically took between 5–10 minutes to complete the surveys. Median completing times were 5.45 min. for WS16, 7.12 min. for ICCSS17, 5.18 min for ECSS18, and 8.17 min. for ECIR19. The

second surveys conducted at ECSS18 and ECIR19 were shorter, with median completion times of 0.87 and 1.30 min, respectively.

Discussion

The data presented here covers many aspects of social behaviour and individual difference constructs relevant to personality science. Its main advantage is the parallel collection of quantitative sensor data about social interactions and survey data about the individuals involved in these interactions. These rich data allow for an exploration of the linkage between a person's characteristics and their social behaviour as measured by the sensors. Furthermore, we present not only one but four data sets collected using the same protocol, making it possible to check for the replicability and reproducibility of results across events. Although contacts as collected by the sensors do not strictly correspond to the sociological definition of an interaction, they are a very good proxy for the analysis of human behaviour and provide data with a high spatial and temporal resolution, with less measurement errors and biases.

This said, the available data has some limitations. First, the contact data covers only the conference venues; information about interactions between participants during social events outside the venue would be immensely valuable, but due to both technical and privacy reasons, such data could not be collected. Within the venue, the collected data is also not immune to noise due to the mishandling of sensors, flickering of the signal, etc. Furthermore, some interactions may be missing due to small gaps in the coverage. Second, the response rates to surveys are high but never 100%; since we did not collect any information about the participants which did not answer, we cannot evaluate whether the non-responders share some characteristics or if the studied population (for which we have survey data) is fully representative of the whole crowd of participants. Though the survey data covers a broad range of information about the participants, some axes which could be very relevant are missing: in particular, we did not collect any data about pre-existing relationships between participants, nor about academic ties such as co-authorship or collaborations. Analysing the effect of such pre-existing ties on the participants' behaviour is thus not possible. Finally, we focused only on close, physical interactions as mediated by contacts. Participants most likely also interacted via electronic means, such as electronic communication (phone, texts, emails) or online social networks. Such data was also not collected and thus the comparison between offline and online behaviour cannot be addressed.

Nonetheless, the available data allow for many diverse and interesting investigations. As a study of different scientific crowds, it first and foremost is valuable for the sociology of science, allowing to determine how different attributes of individuals within such a crowd—status, background, gender—influence their position in the network of interactions. Second, in a more general approach one can consider these setups as examples

of a typical crowd, and investigate the relations between attributes and behaviour. In particular, the availability of information about personality and motivation allow for a comparison between different hypotheses about the predictors of behaviour. Third, in a network science/sociophysics perspective one can look for common points in behaviour in order to investigate potential general mechanisms in the functioning of an assembly of social individuals. Finally, the perception gap experiments open a window on questions such as the effect of the social structure on an individual's perception of the composition of a population.

Among the many possible research questions that can be addressed with this data, we are currently working on two. First, we are exploring the relationship between sociodemographic characteristics and social interactions. We aim to establish whether different sociodemographic groups exhibit consistent variation in the number of connections they establish and their intensity. Second, we investigate the predictive power of personality traits as defined by the Big Five model for the social behaviour participants exhibit at the conferences. Yet, these studies use only a fraction of the data's potential. Therefore, we invite other personality scientists to make use these data to explore individual differences in social behaviour as well as to pinpoint their determinants and correlates.

Funding: M.G. acknowledges support from the Agence Nationale de la Recherche (ANR) project DATAREDUX (ANR19-CE46-0008).

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Author Contributions: *Mathieu Génois*—Idea, conceptualization | Design planning | Resource provision (materials, participants, etc.) | Research implementation (software, hardware, etc.) | Data collection | Data management (storage, curation, processing, etc.) | Visualization (data presentation, figures, etc.) | Data analysis | Validation, reproduction, checking | Writing | Feedback, revisions | Project coordination, administration. *Maria Zens*—Design planning | Resource provision (materials, participants, etc.) | Data collection | Validation, reproduction, checking | Writing | Feedback, revisions | Project coordination, administration. *Marcos Oliveira*—Resource provision (materials, participants, etc.) | Data collection | Writing | Feedback, revisions. *Clemens M. Lechner*—Design planning | Resource provision (materials, participants, etc.) | Research implementation (software, hardware, etc.) | Data collection | Data management (storage, curation, processing, etc.) | Data analysis | Validation, reproduction, checking | Writing | Feedback, revisions | Project coordination, administration. *Johann Schaible*—Resource provision (materials, participants, etc.) | Data collection | Writing | Feedback, revisions | Project coordination, administration. *Markus Strohmaier*—Idea, conceptualization | Writing | Feedback, revisions.

Ethics Statement: Forms of consent for the collection of data about participants to the conferences were validated by legal experts at GESIS. Compliance officer at GESIS confirmed that the hosted data complies with GDPR regulations. Principles on the management of the data collected during these studies and on the restrictions put in place to limit their misuse were agreed upon with the Data Protection Officer of Aix-Marseille Univ.

Other Manuscript Versions: A preprint version of the manuscript before reviews is available on arXiv.org at the following link: https://arxiv.org/abs/2206.05201

Data Availability: Contact data and excerpts of the metadata are available after registration to GESIS's SowiDataNet|datorium. Complete metadata are available upon request to Mathieu Génois (mathieu.genois@cpt.univ-mrs.fr). Data use and access is restricted to scientific purposes only. See the Data section for more details.

Supplementary Materials

For this article, the following Supplementary Materials are available (for access see Index of Supplementary Materials below):

- Venue plans and example of form of consent
- Script to produce Figures 3 and 4
- Open peer-review

Index of Supplementary Materials

Génois, M., Zens, M., Oliveira, M., Lechner, C. M., Schaible, J., & Strohmaier, M. (2023a). *Supplementary materials to "Combining sensors and surveys to study social interactions: A case of four science conferences"* [Venue plans, form of consent]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.12865

Génois, M., Zens, M., Oliveira, M., Lechner, C. M., Schaible, J., & Strohmaier, M. (2023b). *Supplementary materials to "Combining sensors and surveys to study social interactions: A case of four science conferences"* [Script to produce Figures 3 and 4]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.12866

Personality Science. (Ed.). (2023). *Supplementary materials to "Combining sensors and surveys to study social interactions: A case of four science conferences"* [Open peer-review]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.12867

References

APA. (2022). *Dictionary of Psychology*. https://dictionary.apa.org/social-behavior.

Back, M. D. (2021). Chapter 8 - Social interaction processes and personality. In J. F. Rauthmann (Ed.), *The Handbook of personality dynamics and processes* (pp. 183–226). Academic Press. https://doi.org/10.1016/B978-0-12-813995-0.00008-X

Breil, S. M., Geukes, K., Wilson, R. E., Nestler, S., Vazire, S., Back, M. D., & Donnellan, M. B. (2019). Zooming into real-life extraversion–How personality and situation shape sociability in social interactions. *Collabra: Psychology, 5*(1), Article 7. https://doi.org/10.1525/collabra.170

Calabrese, F., Ferrari, L., & Blondel, V. D. (2015). Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys, 47*(2), Article 25. https://doi.org/10.1145/2655691

Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J., & Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLOS ONE, 5*(7), Article e11596. https://doi.org/10.1371/journal.pone.0011596

Choudhury, T., & Pentland, A. (2003). Sensing and modeling human networks using the sociometer. *Proceedings of the Seventh IEEE International Symposium on Wearable Computers* (pp. 216–222). IEEE. https://www.cs.cornell.edu/~tanzeem/pubs/choudhury_iswc2003.pdf

Elmer, T., Chaitanya, K., Purwar, P., Stadtfeld, C. (2019). The validity of RFID badges measuring face-to-face interactions. *Behavior Research Methods, 51*, 2120–2138. https://doi.org/10.3758/s13428-018-1180-y

Génois, M., & Barrat, A. (2018). Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science 7*, Article 11. https://doi.org/10.1140/epjds/s13688-018-0140-1

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John & R. W. Robins (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). The Guilford Press.

Kiti, M. C., Tizzoni, M., Kinyanjui, T. M., Koech, D. C., Munywoki, P. K., Meriac, M., Cappa, L., Panisson, A., Barrat, A., Cattuto, C., & Nokes, D. J. (2016). Quantifying social contacts in a

household setting of rural Kenya using wearable proximity sensors. *EPJ Data Science, 5*, Article 21. https://doi.org/10.1140/epjds/s13688-016-0084-2

Kontro, I., & Génois, M. (2020). Combining surveys and sensors to explore student behaviour. *Education Sciences, 10*(3), Article 68. https://doi.org/10.3390/educsci10030068

Malik, M. M. (2018). *Bias and beyond in digital trace data* [Doctoral dissertation]. Carnegie Mellon University. http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

Mastrandrea, R., Fournet, J., Barrat., A. (2015). Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE, 10*(9), Article e0136497. https://doi.org/10.1371/journal.pone.0136497

Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics, 4*(3). https://doi.org/10.1177/2053168017720008

Oliveira, M., Karimi, F., Zens, M., Schaible, J., Génois, M., & Strohmaier, M. (2022). Group mixing drives inequality in face-to-face gatherings. *Communications Physics, 5*, Article 127. https://doi.org/10.1038/s42005-022-00896-1

Ozella, L., Paolotti, D., Lichand, G., Rodríguez, J. P., Haenni, S., Phuka, J., Leal-Neto, O. B., & Cattuto, C. (2021). Using wearable proximity sensors to characterize social contact patterns in a village of rural Malawi. *EPJ Data Science, 10*, Article 46. https://doi.org/10.1140/epjds/s13688-021-00302-w

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology, 107*(4), 677–718. https://doi.org/10.1037/a0037250

Rauthmann, J. F., & Sherman, R. A. (2016). Ultra-brief measures for the situational eight DIAMONDS domains. *European Journal of Psychological Assessment, 32*(2), 165–174. https://doi.org/10.1027/1015-5759/a000245

Rojas, M. B., Sadeghvaziri, E., & Jin, X. (2016). Comprehensive review of travel behavior and mobility pattern studies that used mobile phone data. *Transportation Research Record, 2563*(1), 71–79. https://doi.org/10.3141/2563-11

Rout, A., Nitoslawski, S., Ladle, A., & Galpern, P. (2021). Using smartphone-GPS data to understand pedestrian-scale behavior in urban settings: A review of themes and approaches. *Computers, Environment and Urban Systems, 90*, Article 101705. https://doi.org/10.1016/j.compenvurbsys.2021.101705

Salathé, M., Kazandjieva, M., Lee, J. W., Levis, P., Feldman, M. W., & Jones, J. H. (2010). A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences, 107*(51), 22020–22025. https://doi.org/10.1073/pnas.1009094108

Schaible, J., Oliveira, M., Zens, M., & Génois, M. (2022). Sensing close-range proximity for studying face-to-face interaction. In U. Engel, A. Quan-Haase, S. Liu, & L. Lyberg (Eds.), *Handbook of computational social science, Volume 1 - Theory, case studies and ethics* (pp. 215–236). Routledge.

Sila-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science, 30*(5), 881–906. https://doi.org/10.1080/13658816.2015.1100731

Soto, C. J. & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality, 68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004

Stehlé, J., Charbonnier, F., Picard, T., Cattuto, C., Barrat., A. (2013). Gender homophily from spatial behavior in a primary school: A sociometric study. *Social Networks, 35*(4), 604–613. https://doi.org/10.1016/j.socnet.2013.08.003

Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., & Lehmann, S. (2014). Measuring large-scale social networks with high resolution. *PLOS ONE, 9*(4), Article e95978. https://doi.org/10.1371/journal.pone.0095978

Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Régis, C., Kim, B.-a., Comte, B., & Voirin, N. (2013). Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLOS ONE, 8*(9), Article e73970. https://doi.org/10.1371/journal.pone.0073970

Vu, L., Nahrstedt, K., Retika, S., & Gupta, I. (2010). Joint Bluetooth/Wifi scanning framework for characterizing and leveraging people movement in university campus. In *Proceedings of the 13th ACM International Conference on Modeling, Analysis, and Simulation of Wireless and Mobile Systems, MSWIM '10* (pp. 257–265). https://doi.org/10.1145/1868521.1868563

eapp

Personality Science (PS) is an official journal of the European Association of Personality Psychology (EAPP).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.