Measuring contact patterns with wearable sensors: methods, data characteristics and applications to data-driven simulations of infectious diseases

A. Barrat[1,2,3,§], C. Cattuto[3], A.E. Tozzi[4], P. Vanhems[5,6], N. Voirin[5]

[1]Aix Marseille Université, CNRS, CPT, UMR 7332, 13288 Marseille, France

[2]Université de Toulon, CNRS, CPT, UMR 7332, 83957 La Garde, France

[3]Data Science Laboratory, ISI Foundation, Torino, Italy

[4]Bambino Gesù Children's Hospital, IRCCS, Rome, Italy

[5]Hospices Civils de Lyon, Hôpital Edouard Herriot, Service d'Hygiène, Epidémiologie et Prévention, Equipe Epidémiologie et Biomarqueurs de l'Infection, Lyon, France

[6]Université de Lyon; université Lyon 1; CNRS UMR 5558, laboratoire de Biométrie et de Biologie Evolutive, Equipe Epidémiologie et Santé Publique, Lyon, France

[§]Corresponding author

Abstract

Thanks to recent technological advances, measuring real-world interactions using mobile devices and wearable sensors has become possible, allowing researchers to gather data on human social interactions in a variety of contexts with a high spatial and temporal resolution. Empirical data describing contact networks reach thus a high level of detail that may yield bring new insights into the dynamics of infection transmission between individuals. At the same time, such data bring forth new challenges related to their statistical description and analysis and to their use in mathematical models. In particular, the integration of highly detailed empirical data in computational frameworks designed to model the spread of infectious diseases raises the issue of assessing which representations of the raw data work best to inform the models. There is an emerging need to strike a balance between simplicity and detail in order to ensure both generalizability and accuracy of predictions. Here, we review recent work on the collection and analysis of highly detailed data on temporal networks of face-to-face human proximity, carried out in the context of the SocioPatterns collaboration. We discuss the various levels of coarse-graining that can be used to represent the data in order to inform models of infectious diseases transmission. We moreover discuss several limitations of the data and future avenues for data collection and modeling efforts in the field of infectious diseases.

Introduction

Contact patterns among individuals play an important role in determining the potential transmission routes of infectious diseases. Knowledge of these patterns is thus relevant for identifying contagion pathways, for informing models of epidemic spread, and for the design and evaluation of control measures such as the targeting of specific groups of individuals with appropriate prevention strategies or interventions (e.g., drug prophylaxis, vaccination, hand washing, use of masks...) [1,2,3,4,5].

Empirical descriptions of contact patterns have until recently mostly relied on interviews and surveys, sometimes at very large scale [6,7,8,9,10,11,12,13], yielding important insights. Surveys allow distinguishing between different types of contacts (e.g., involving or not physical contact) and to classify the contacts according to their context (at work, at home, at school,…), and have yielded information on the mixing patterns of different age groups in various countries [10].

Surveys are however costly, have often a low response rate [13], and the precise formulation of the question might influence the answers. Answers are also subject to memory biases, which are difficult to estimate [12]. Moreover, surveys often collect ego-networks on single days (see however [12,14]), and it is then difficult to estimate some properties of the contact networks known to be relevant for the spread of infectious diseases, such as the number of triangles and the fraction of repeated contacts from one day to the next [15,16].

The use of novel technologies, in particular of networked wearable sensors, offers appealing alternatives: Bluetooth or Wi-Fi can be used to assess proximity of individuals [17,18], and even face-to-face presence of individuals can be resolved with high spatial and temporal resolution [19]. Here, we review recent work developed within the SocioPatterns collaboration (www.sociopatterns.org), where wearable proximity sensors were used to collect large-scale datasets on human face-to-face interactions in various contexts, including conferences, hospitals, schools, and museums [19,20,21,16].

Data collection: method, statistical analysis, and representations

The data collection infrastructure is based on wearable wireless devices that exchange radio packets in a peer-to-peer fashion to monitor location and proximity of individuals (www.sociopatterns.org). The use of ultra-low power radio signals allows radio packets to be exchanged only between devices located within 1–1.5 meters of one another. Moreover, when individuals wear the devices

on their chest and the lowest radio power is used, exchange of packets between devices is only possible when they are facing each other, as the human body acts as a shield at the radio frequency used. In summary, the system detects and records close-range encounters during which a communicable disease infection could be transmitted, for example, by coughing, sneezing or hand contact (see Figure 1 and [19]).

The sensing system is tuned so that the recorded data include, for each detected contact between participants, its starting and ending times, with a temporal resolution of about 20 seconds: it is thus possible to monitor the number of contacts that each individual establishes with any other individual, the duration of individual encounters, the cumulative time spent in contact between two or more individuals, the frequency of encounters, and how these quantities evolve during the observation period.

The complexity of the data is exposed through the statistical analysis of the contact event durations and of the time intervals between contacts: a large variability is indeed observed in all these quantities. The corresponding distributions, shown in Fig. 2A) and B), are broad, similarly to other characteristics of human behavior [22]: short durations are the most probable, but very long durations are also observed with a non-negligible probability, and no characteristic temporal scale emerges. For transmissible diseases for which the transmission probability between two individuals depends on their time in contact, this means that different contacts might yield very different transmission probabilities: many contacts are very short and correspond to a small transmission probability, but some are much longer than others and could therefore play a crucial role in disease dynamics, These strong duration fluctuations are a robust property observed across all contexts [19,20,21] and at different moments despite the amount of activity varies significantly from case to case (Fig. 2C).

The detailed knowledge of all contact events allows to recreate an artificial in silico population and hence to simulate potential spreading phenomena with a high degree of realism. It is also possible to test, in simulation, the impact of specific interventions for mitigation and containment. Highly detailed contact data might however represent unnecessarily detailed information if the goal is to simply extract stylized facts or generic statistics of contacts, which can be used to design and validate models of human contacts and to estimate the relevance of transmission control strategies. The observed sequence of contacts might indeed be influenced by specific aspects of the environment in which the measure took place, and only represents one instance of many contact

sequences taking place in the same environment in different days. It is thus useful to build contact summary statistics, which are expected to be more robust to variations of the specific measurement context. To this aim, the time-resolved contact data are usually aggregated along two dimensions, as discussed below.

First, aggregation along the temporal dimension yields cumulative contact networks preserving the information at the individual level: such networks describe who has been in contact with whom and each link between two nodes is weighted by the cumulative time spent in contact by the two corresponding individuals. The resulting weights are highly heterogeneous, as shown in Fig 2D): while most links correspond to very short durations, some correspond to very long contact times, with no characteristic interaction timescale. Similar statistics have been observed in various contexts ranging from scientific conferences to schools, hospitals or offices [19,16,3,20,21,23]. The heterogeneity of contact patterns at the individual level is known to have a strong impact on spreading dynamics [24,25]. In particular, it highlights the existence of "super-contactors", i.e., individuals who account for an important part of the overall contact durations and may therefore become super-spreaders in case of an outbreak [26,23].

High-resolution contact data can also be aggregated over specific attributes of the individuals. When the population under study is structured, i.e., when individuals can be classified according to specific characteristics or role (e.g., according to their age class or professional activity), a convenient representation of their contacts is provided by contact matrices whose elements give the number (or duration) of the contacts that individuals in one given class have with individuals of another class, as illustrated in Fig. 3.. Such a representation is useful for designing interventions as it can suggest easily generalizable strategies that target specific classes of individuals.

Contact matrices, however, typically report only averages, discarding the strong fluctuations in the numbers and durations of contacts between two individuals of given classes. The contact matrix representation also carries the implicit assumption that all individuals are in contact with one another: any two individuals are assumed to be connected, with a weight that only depends on their relative classes.

The cumulative duration of contacts between two individuals, however, fluctuates strongly, even when their role classes are fixed [21,27]: for instance, the contact duration between nurse A and doctor B can be very different from the one between nurse C and doctor D. The average contact duration between nurses and doctors is thus not a sufficient information. Moreover, the density of

links connecting individuals in given classes depends on the specific classes and is sometimes very small: many pairs of individuals never have any contact. In order to account for these important properties, the concept of "contact matrix of distributions" (CMD) was introduced [27]: in this representation, as in usual contact matrices, the contact patterns between individuals depend on their relative classes. For each pair of classes, the distribution of cumulated contact durations is fitted to a certain functional form (e.g., a negative binomial in order to account for its broad character) and the matrix elements are given by the parameters of this fit. The CMD representation therefore does not retain the specific information on who has been in contact with whom but it does retain both the empirical density of links and the heterogeneity of contact durations.

Finally, each of the above representations can be computed for the entire duration spanned by the dataset or for restricted time windows, such as half-days or days, in order to investigate possible variations with time of the contact statistics, networks or matrices with time [20,23,27]. Information on temporal variations of individuals contacts from one day to the next could be included in the contact matrix data summaries, for instance as an additional parameter in each contact matrix entry (giving, e.g. for nurses and doctors, the average fraction of new contacts between nurses and doctors from one day to the next).

Using high-resolution contact data in models of epidemic spread

Empirical time-resolved contact data can be used to perform data-driven simulations of epidemic spread in a population [16,27,3]. As each dataset describes a specific population and environment, and therefore features sprecificities that might not be representative of another period or context, an issue naturally emerges: what level of detail on the contact patterns should be incorporated into computational models of spread, so that the relevant information is retained but the model stays as parsimonious and generalizable as possible? In other words, what are the most useful synopses of high-resolution contact patterns? On the one hand, too coarse data summaries might disregard important properties; on the other hand, too fine representations might be too specific and the integration of highly detailed data may yield models that are less transparent and lead to results that are less general in their applicability [16,27,28].

In order to shed light on these issues, we have investigated the impact of the data representation on simulations of epidemic spread by building a hierarchy of data representations corresponding to different levels of aggregation. Each representation was used to simulate the spread of an infectious disease [16,27] and the results were compared with the outcome of simulations based on the most

detailed representation (regarded as a gold standard). This methodology provides several insights: First, the epidemic peak timing is a robust property of the spread. It is correctly approximated even by simulations based on coarse data representations, at least for spreading processes which are slow with respect to the data temporal resolution [16,27]. Moreover, data representations that do not take into account the heterogeneity of contact durations lead to an overestimation of the probability of a large outbreak and of the attack rate (see Fig. 4) [16,27]. Most importantly, they might lead to an incorrect classification of the relative risks for different classes of individuals [27]. Compared to this, the representation by contact matrices of distributions allows to correctly model important features of epidemic spread and to estimate the relative risks of individuals in different role classes, while maintaining a parsimonious form that retains very little information from the time-varying contact data it summarizes. This representation thus provides a practical tool that translates complex properties of the contacts within a population into practically actionable information for guiding intervention and prevention strategies. It represents an interesting synopsis of highly detailed contact data that combines simplicity, compactness of representation and modeling power, which are essential features for guiding decision making in public health contexts.

Discussion

Infection control still represents one of the most important challenges in public health. The current policies for control of infections transmitted through person to person contact are based on general assumptions that may not be applicable to all individuals and that may be difficult to deliver to an entire population. Strategies driven by detailed data on the contact patterns within populations promise to increase efficiency and feasibility. To this end, it is crucial to identify appropriate methods for using high-resolution data to inform models and design prevention strategies.

The unsupervised measurement of contact patterns with wearable sensors provides an interesting opportunity in this direction: It gives access to the network of contacts between individuals, and provides key information on the structure and heterogeneity of contacts between individuals belonging to different role classes and on the repetition of contact patterns in different days. The collection of data in diverse contexts highlights differences and similarities of human contacts depending on context, and has shown the remarkable robustness of crucial statistical features such as the distribution of cumulated contact durations. The collected data can be used to design models of human behavior, inform models of epidemic spread, and design and evaluate containment strategies in diverse contexts such as schools or hospitals. For instance, it is possible to evaluate, in

simulation, the performance of targeted vaccination strategies or the role of the intervention timing. It is also possible to envision novel intervention types such as changes in the daily schedule of a school or in the organization of a hospital ward, or to estimate the relative efficiency of school closure and targeted class closure [29].

The present methodology carries some limitations. Most datasets collected so far correspond to the contacts in populations of relatively limited size (few hundred individuals) over a limited amount of time. No information is collected on contacts occurring either outside the range of the sensing system or involving individuals not wearing sensors. Moreover, the data do not provide information on physical contacts (unlike some surveys) nor on the occurrence of events that are known to favor transmission such as coughing or sneezing.

These limitations hint at future research directions. Further data collection campaigns will be crucial to validate and consolidate the results across other hospital units, other contexts, and over longer periods of time. Detailed comparisons with surveys [13] would also be an important cross-validation tool. The role of sampling of the population should also be carefully assessed.

Additional datasets will also help evaluate the use of proxies, such as the ones put forward in [2,30], that may replace systematic detailed measurement of contact patterns. For example, using metadata on school schedules or shifts in hospitals allows to infer approximate colocation properties, group structures and spatial distributions that can be used to simulate epidemic spread and design intervention strategies. High-resolution data from wearable sensors could then be used as a gold standard to validate the results of the simulations that use only proxy data.

High-resolution datasets can also be used to devise models of human mobility and contact used to generate synthetic datasets fed into models of disease spread at various scales, helping to eventually achieve multi-scale models that span several temporal, spatial and contact scales.

The overlay and integration of epidemiological, microbiological and genomic information with contact patterns among individuals may moreover enable radical changes in the approach to infection control. The precise measure of the contact pattern in a population could be combined with whole genome sequencing techniques of pathogens to investigate outbreaks. So far only traditional surveys have been used to investigate social networks in these situations [31]. Combining phylogenetic analysis of viruses with actual contact data could provide valuable information about the transmission mechanisms of infectious diseases, especially regarding the role

of frequency and/or duration of contacts. The inclusion of host susceptibility characteristics (e.g., age, sex, underlying diseases, genetic susceptibility, etc) in such datasets would also enable more precise studies of infectious diseases transmission mechanisms.

Acknowledgments

This work has been partially supported by the French ANR project HarMS-flu (ANR-12-MONU-0018) to AB, and by the EU FET project Multiplex 317532 to AB and CC. This study was partially supported by the FINOVI foundation, the INSERM IMI Pandemic program and GOJO.

References

1. Chowel G, Viboud C (2013) A practical method to target individuals for outbreak detection and control. BMC Med 11:36.
2. Smieszek T, Salathe M (2013) A low-cost method to assess the epidemiological importance of individuals in controlling infectious disease outbreaks. BMC Med 11: 35.
3. Salathe M, Kazandjieva M, Lee JW, Levis P, Feldman MW, et al. (2010) A high-resolution human contact network for infectious disease transmission. Proc Natl Acad Sci U S A 107: 22020-22025.
4. Cauchemez S, Bhattarai A, Marchbanks TL, Fagan RP, Ostroff S, Ferguson NM, Swerdlow D (2011) Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. Proc Natl Acad Sci USA 108:2825–2830.
5. Temime L, Opatowski L, Pannet Y, Brun-Buisson C, Boelle PY, et al. (2009) Peripatetic health-care workers as potential superspreaders. Proc Natl Acad Sci U S A 106: 18420-18425.
6. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DA (2012) Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiol Infect 140: 2117-2130.
7. Beutels P, Shkedy Z, Aerts M, Van Damme P (2006) Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface. Epidemiol Infect 134: 1158-1166.
8. McCaw JM, Forbes K, Nathan PM, Pattison PE, Robins GL, et al. (2010) Comparison of three methods for ascertainment of contact information relevant to respiratory pathogen transmission in encounter networks. BMC Infect Dis 10: 166.
9. Mikolajczyk RT, Akmatov MK, Rastin S, Kretzschmar M (2008) Social contacts of school children and the transmission of respiratory-spread pathogens. Epidemiol Infect 136: 813-822.
10. Mossong J, Hens N, Jit M, Beutels P, Auranen K, et al. (2008) Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 5: e74.
11. Zagheni E, Billari FC, Manfredi P, Melegaro A, Mossong J, et al. (2008) Using time-use data to parameterize models for the spread of close-contact infectious diseases. Am J Epidemiol 168: 1082-1090.
12. Smieszek T, Burri EU, Scherzinger R, Scholz RW (2012) Collecting close-contact social mixing data with contact diaries: reporting errors and biases. Epidemiol Infect 140: 744-752.

13. Danon L, Read JM, House TA, Vernon MC, Keeling MJ (2013) Social encounter networks: characterizing Great Britain. Proc. Roy. Soc. B 280:20131037.
14. Read JM, Eames KT, Edmunds WJ (2008) Dynamic social networks and the implications for the spread of infectious disease. J. Royal Soc. Interface; 5: 1001–1007.
15. Smieszek T, Fiebig L, Scholz RW (2009) Models of epidemics: when contact repetition and clustering should be included. Theor Biol Med Model 6: 11.
16. Stehlé J, Voirin N, Barrat A, Cattuto C, Colizza V, et al. (2011) Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. BMC Med 9: 87.
17. O'Neill E, Kostakos V, Kindberg T, Fatah gen. Schieck A, Penn A, et al (2006) Instrumenting the city: developing methods for observing and understanding the digital cityscape. Lecture Notes in Computer Science 4206:315-332.
18. Pentland A (2008) Honest Signals: how they shape our world. MIT Press, Cambridge MA.
19. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, et al. (2010) Dynamics of person-to-person interactions from distributed RFID sensor networks. PLoS One 5: e11596.
20. Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, et al. (2011) High-resolution measurements of face-to-face contact patterns in a primary school. PLoS One 6: e23176.
21. Isella L, Romano M, Barrat A, Cattuto C, Colizza V, et al. (2011) Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors. PLoS One 6: e17144.
22. Barabasi A-L (2010). Bursts: The Hidden Pattern Behind Everything We Do. Dutton Adult.
23. Vanhems P, Barrat A, Cattuto C, Pinton J-F, Khanafer N, Régis C, Kim B, Comte B, Voirin N (2013) Estimating potential infection transmission routes in hospital wards using wearable proximity sensors, PloS ONE 8:73970.
24. Anderson, RM, May, RM (1992) Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford.
25. Pastor-Satorras R, Vespignani A (2001), Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86 (2001) 3200-3203.
26. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM (2005) Superspreading and the effect of individual variation on disease emergence. Nature 438: 355-359.
27. Machens A, Gesualdo F, Rizzo C, Tozzi AE, Barrat A, Cattuto C (2013) An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices, BMC Infect. Dis. 13:185
28. Blower S, Go MH (2011) The importance of including dynamic social networks when modeling epidemics of airborne infections: does increasing complexity increase accuracy? BMC Medicine 9:88.
29. Gemmetto V, Barrat A, Cattuto C (2013) Mitigation of infectious diseases at school: targeted class closures vs school closures, to be submitted.
30. Curtis DE, Hlady C, Pemmaraju SV, Polgreen P, Segre AM (2010) Modeling and Estimating the Spatial Distribution of Healthcare Workers. Proceedings of the 1st ACM International Health Informatics Symposium IHI'10: 287-296.
31. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N. Engl. J. Med 364:730.

Name	Date	Venue	Event type	# participants	Duration
SG	Apr-Jul 2009	Science Gallery, Dublin, IE	Exhibition	~30,000	3 months
ESWC09	Jun 2009	ESWC 2009, Crete, GR	Conference	~180	4 days
SFHH	Jun 2009	SFHH, Nice, FR	Conference	~400	2 days
HT09	Jul 2009	ACM Hypertext 2009, Torino, IT	Conference	~120	3 days
PS	Oct 2009	Primary School, Lyon, FR	School	~250	2 days
OBG	Nov 2009	Bambino Gesu Hospital, Rome, IT	Hospital	~100	10 days
HOSP	Dec 2010	Edouard Herriot Hospital, Lyon, FR	Hospital	~100	4 days

TABLE 1: Partial list of the datasets on face-to-face proximity collected by the SocioPatterns collaboration during 2009 and 2010 and discussed in the present paper.

FIGURE CAPTIONS

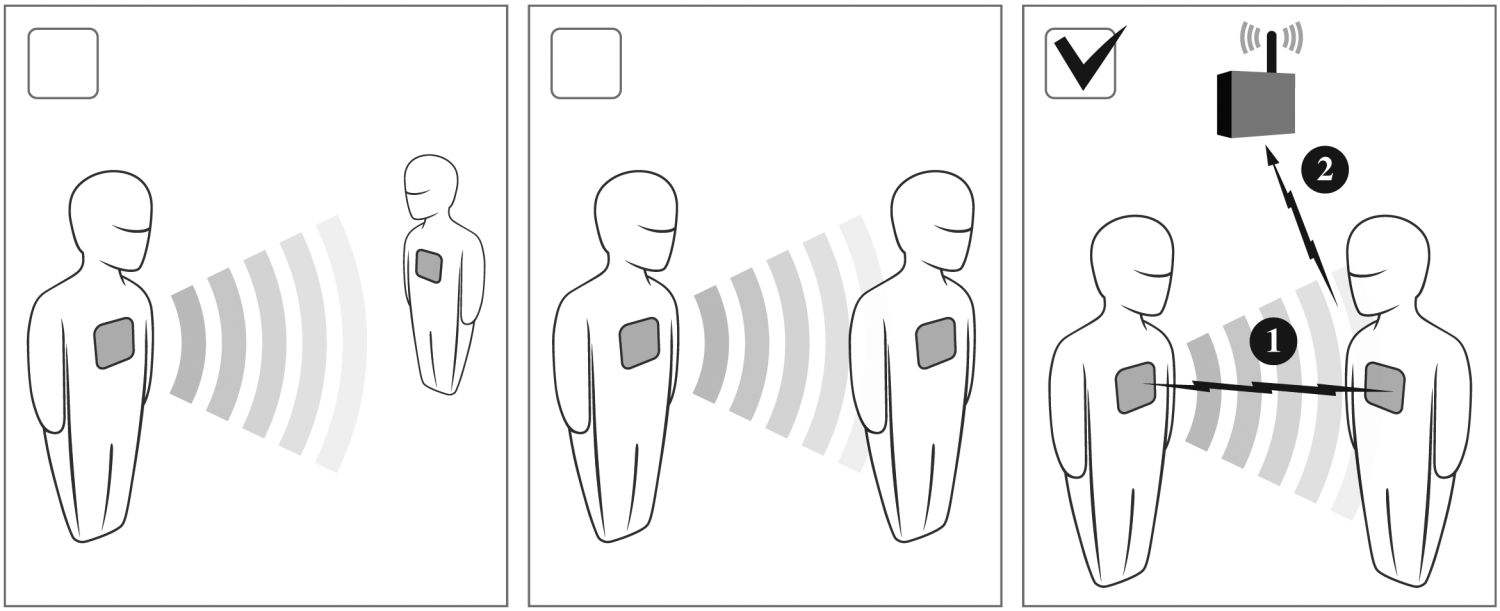
Figure 1: A) Schematic illustration of the sensing infrastructure. RFID devices are worn as badges by the individuals participating to the deployments. A face-to-face contact is detected when two persons are close and facing each other. The interaction signal is then sent to RFID readers located in the environment. B) RFID device worn by participants. C) RFID reader.

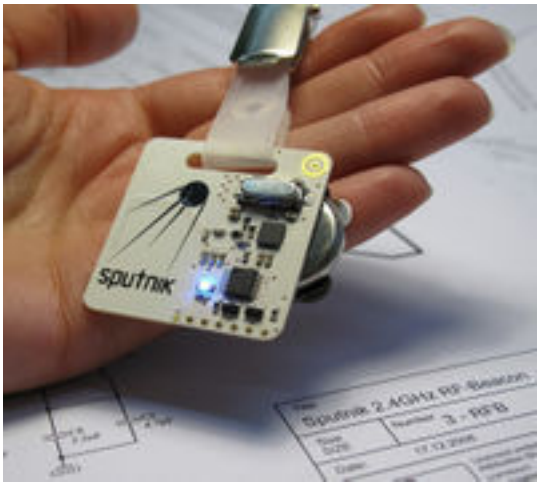
Figure 2: Statistical properties of the contact data for several datasets (see Table for the datasets characteristics). A) Probability of observing a contact of duration Δt vs Δt, computed as the number of contacts of duration Δt divided by the total number of contacts; B) Probability of observing a time interval of a given duration between two successive contact events of a given individual, aggregated over the entire population; C) Evolution of the number of nodes and links in 20-seconds instantaneous networks during a conference; D) Probability of observing a daily cumulated contact duration w_{ij} between individuals i and j (i.e., number of pairs i-j with daily cumulated contact duration w_{ij}, divided by the total number of pairs of individuals who have been in contact at least once during the day).

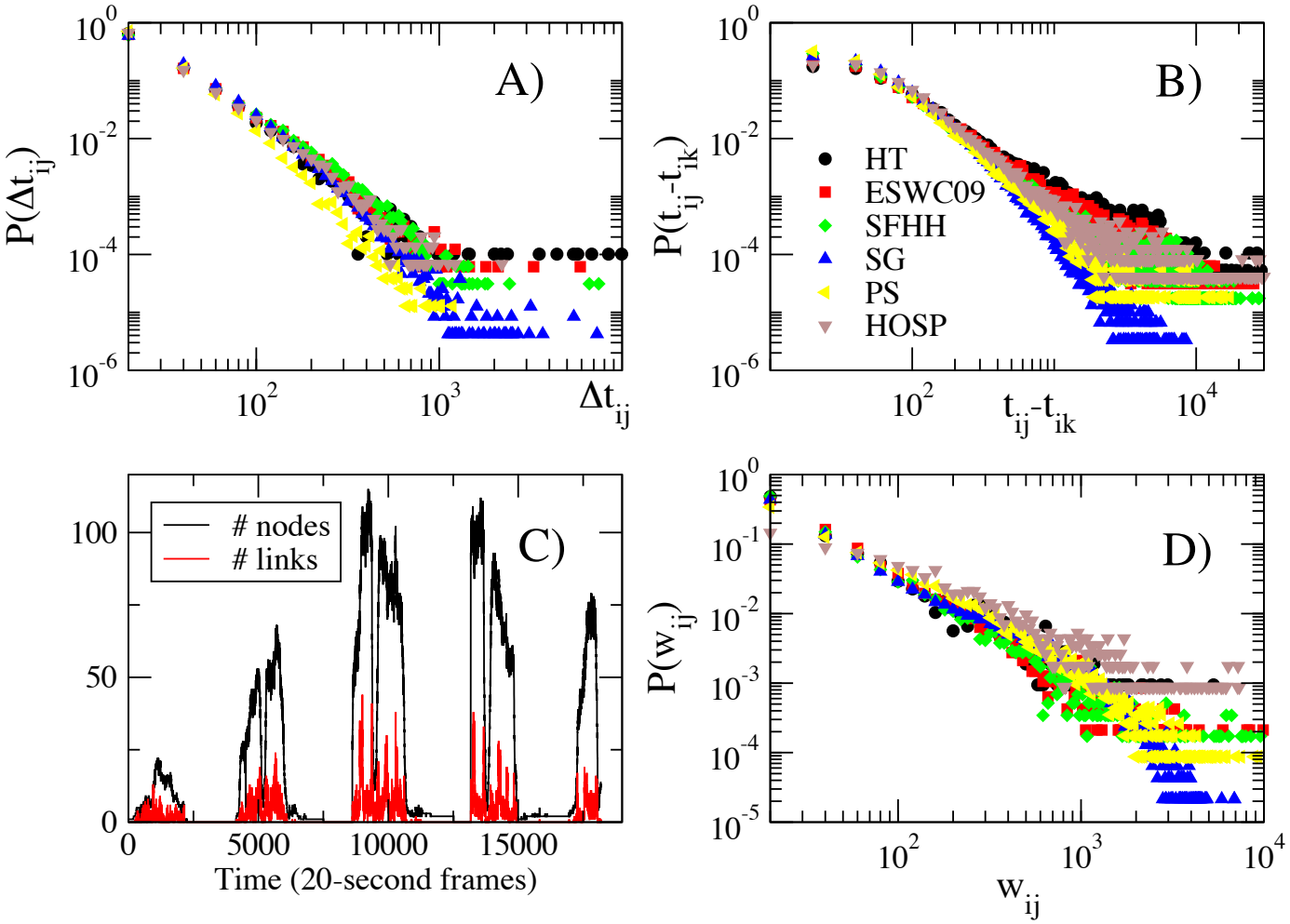
Figure 3: Contact matrices giving the cumulated durations in seconds of the contacts between classes of individuals, measured in the HOSP and PS datasets. In the hospital case, individuals were categorized, according to their role in the ward, as nurses, doctors, patients and administrative staff. In the school, the categorization is given by the division of the students in classes (here ranging from 1^{st} to 5^{th} grade). The matrix entry at row X and column Y gives the total duration of all contacts between all individuals of class X with all individuals of class Y. Abbreviations: NUR, paramedical staff (nurses and nurses' aides); PAT, Patient; MED, Medical doctor; ADM, administrative staff.

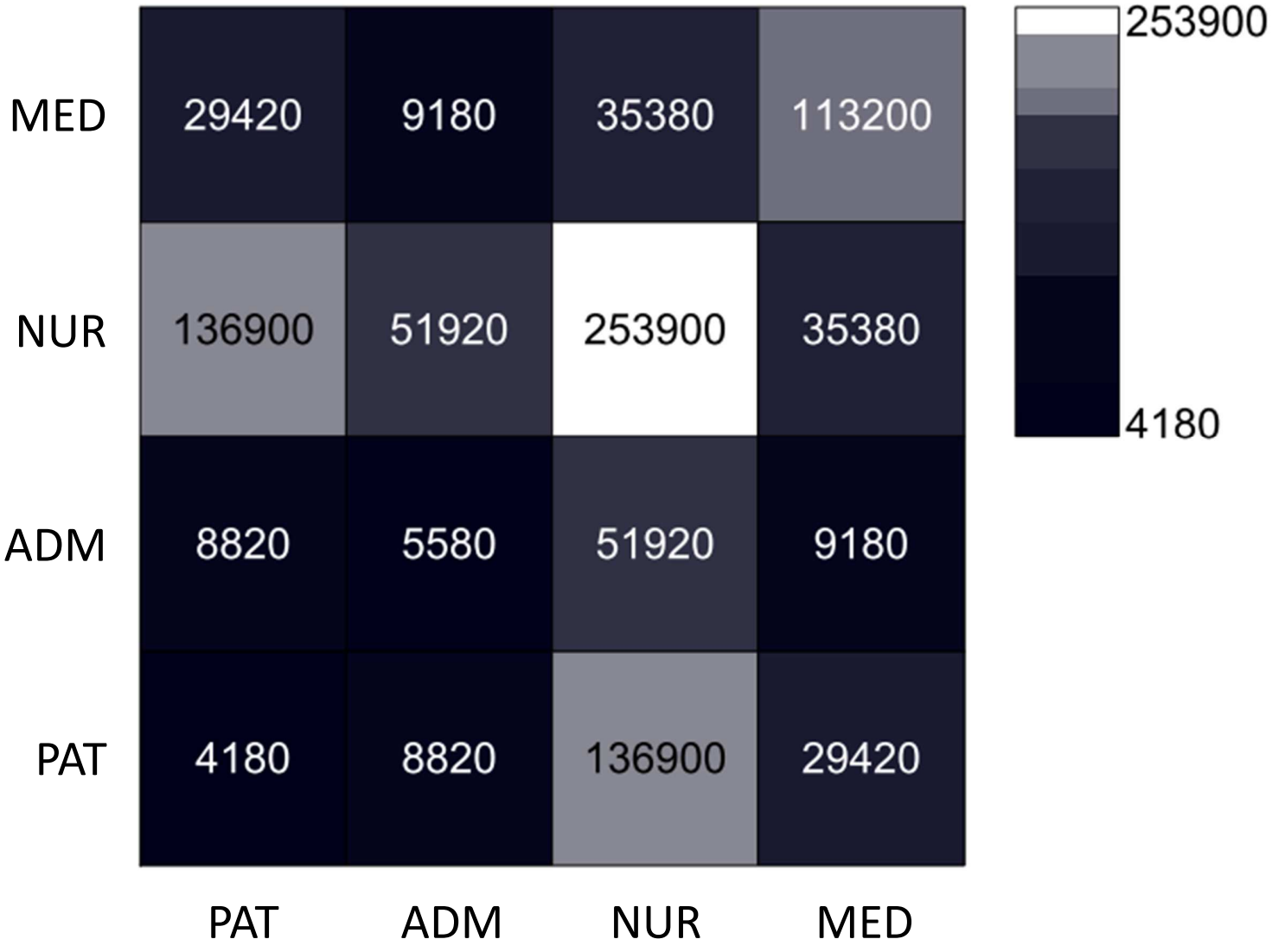
Figure 4: Probability of observing a certain attack rate (fraction of the final number of cases) for numerical simulations of a Susceptible-Exposed-Infected-Recovered model on the OBG dataset (i.e., number of simulations leading to a given attack rate, divided by the total number of simulations). The different curves correspond to simulations performed using different representations of the raw data. DYN: dynamical contact network including the precise starting and ending time of each contact. CM: contact matrix representation that includes only information on the average contact time between individuals of different classes (individuals are here categorized

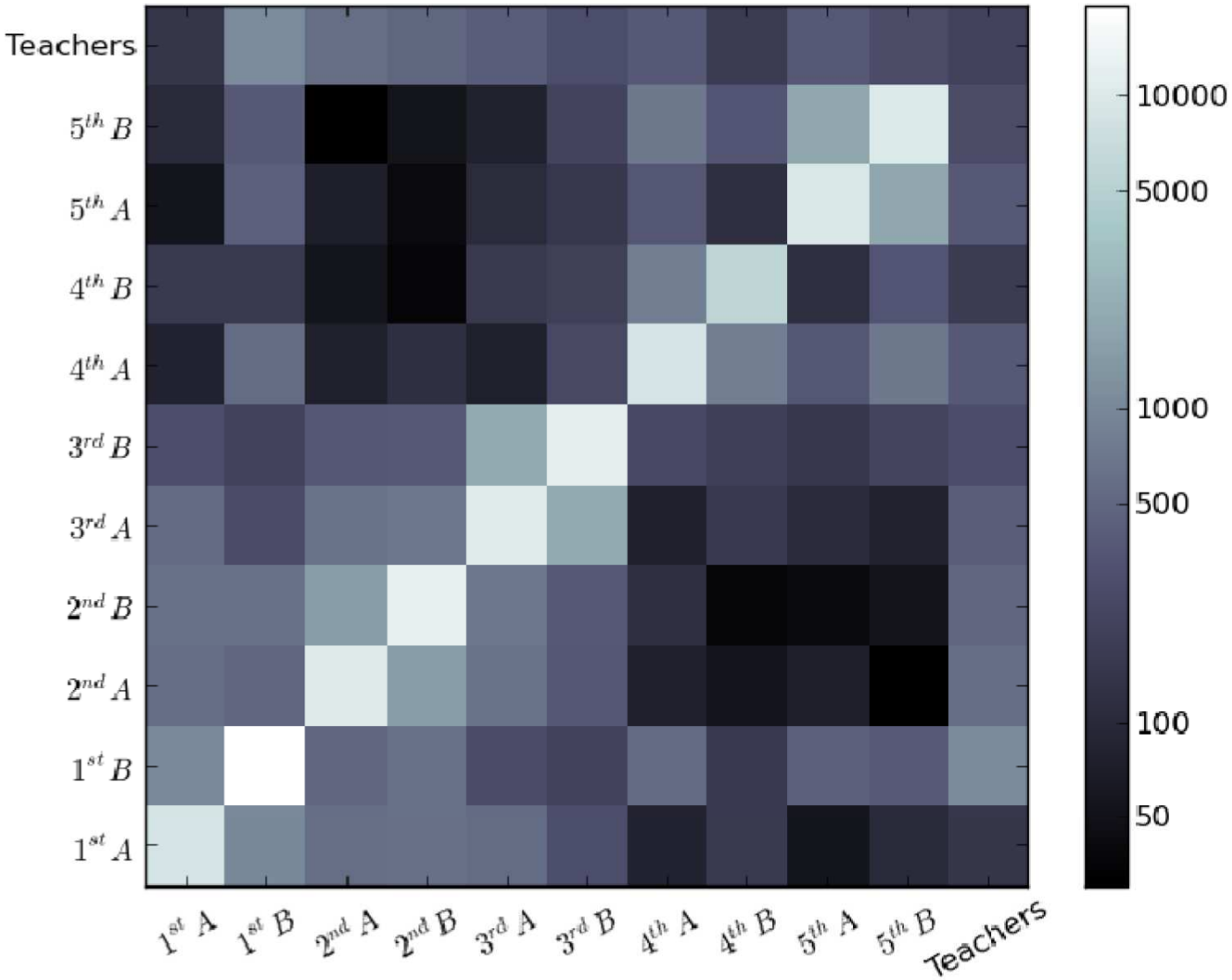
as Nurses, Assistants, Doctors, Patients and Caregivers). CMD: contact matrix of distributions that takes into account the heterogeneity of contact durations and the different density of links among different categories of individuals. Simulations performed using the CM representation lead to an overestimation of the final number of cases.

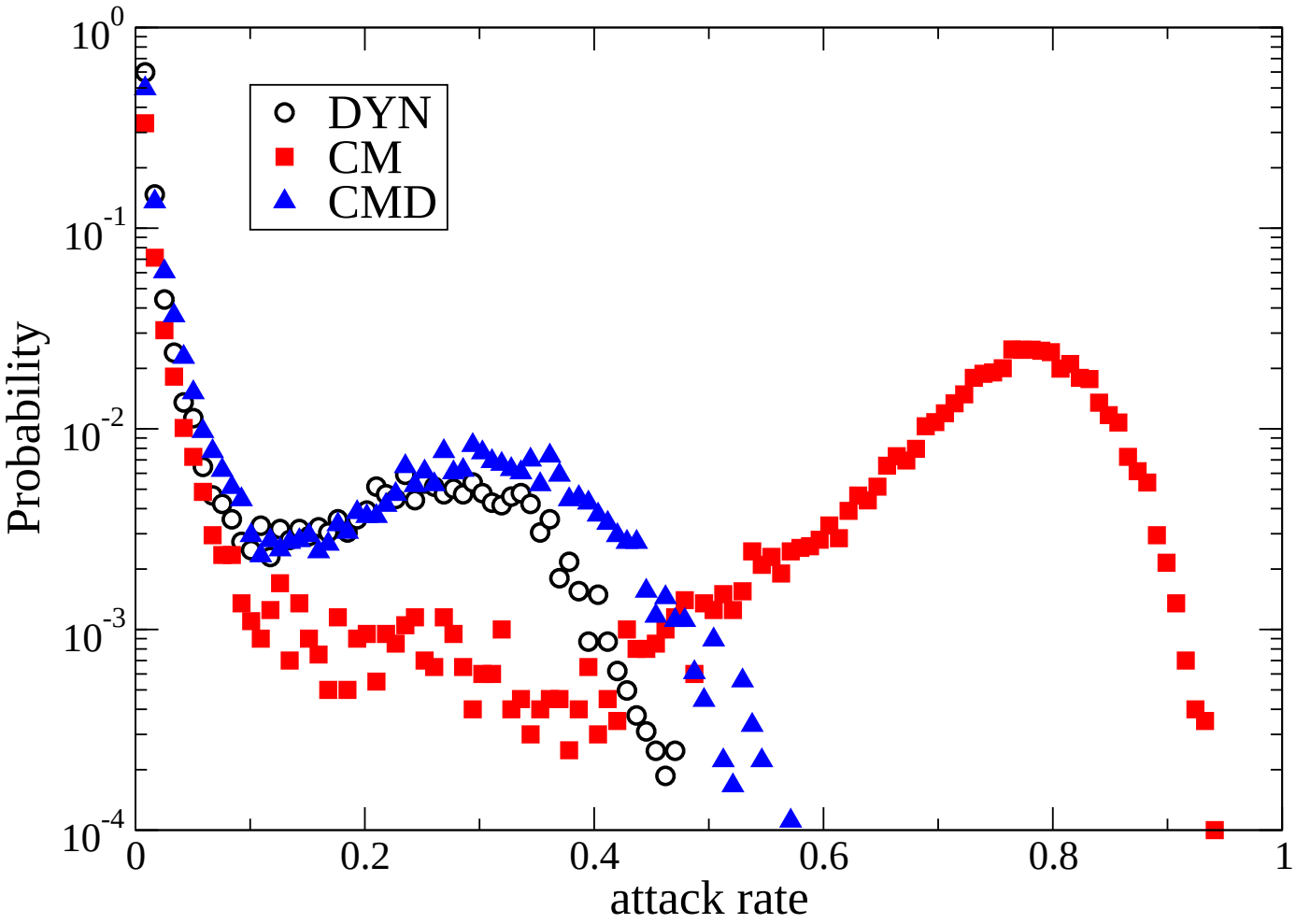

1
2


sputnik


A)
B)
C)
D)
HT
ESWC09
SFHH
SG
PS
HOSP
nodes
links
Time (20-second frames)


MED
NUR
ADM
PAT
29420
9180
35380
113200
136900
51920
253900
35380
8820
5580
51920
9180
4180
8820
136900
29420
PAT
ADM
NUR
MED
253900
4180


Teachers
5^{th} B
5^{th} A
4^{th} B
4^{th} A
3^{rd} B
3^{rd} A
2^{nd} B
2^{nd} A
1^{st} B
1^{st} A
1^{st} A
1^{st} B
2^{nd} A
2^{nd} B
3^{rd} A
3^{rd} B
4^{th} A
4^{th} B
5^{th} A
5^{th} B
Teachers
10000
5000
1000
500
100
50


DYN
CM
CMD
Probability
attack rate
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
0.2
0.4
0.6
0.8
1