Modeling workplace contact networks: The effects of organizational structure, architecture, and reporting errors on epidemic predictions

GAIL E. POTTER, TIMO SMIESZEK and KERSTIN SAILER

Link to this article: http://journals.cambridge.org/abstract_S2050124215000223

How to cite this article:
GAIL E. POTTER, TIMO SMIESZEK and KERSTIN SAILER (2015). Modeling workplace contact networks: The effects of organizational structure, architecture, and reporting errors on epidemic predictions. Network Science, 3, pp 298-325 doi:10.1017/nws.2015.22

Request Permissions : Click here

Modeling workplace contact networks: The effects of organizational structure, architecture, and reporting errors on epidemic predictions

GAIL E. POTTER

California Polytechnic State University, San Luis Obispo, CA, USA
and
Center for Statistics and Quantitative Infectious Disease,
Fred Hutchinson Cancer Research Center, Seattle, WA, USA
(e-mail: `gepotter@calpoly.edu`*)*

TIMO SMIESZEK

Center for Infectious Disease Dynamics, Pennsylvania State University

Modelling and Economics Unit, Centre for Infectious Disease Surveillance and Control, Public Health England, London, UK

MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College School of Public Health, London, UK
and
NIHR Health Protection Research Unit in Modelling Methodology,
Department of Infectious Disease Epidemiology, Imperial College School of Public Health, London, UK
(e-mail: `timo.smieszek@phe.gov.uk`*)*

KERSTIN SAILER

The Bartlett School of Graduate Studies, University College London
(e-mail: `k.sailer@ucl.ac.uk`*)*

Abstract

Face-to-face social contacts are potentially important transmission routes for acute respiratory infections, and understanding the contact network can improve our ability to predict, contain, and control epidemics. Although workplaces are important settings for infectious disease transmission, few studies have collected workplace contact data and estimated workplace contact networks. We use contact diaries, architectural distance measures, and institutional structures to estimate social contact networks within a Swiss research institute. Some contact reports were inconsistent, indicating reporting errors. We adjust for this with a latent variable model, jointly estimating the true (unobserved) network of contacts and duration-specific reporting probabilities. We find that contact probability decreases with distance, and that research group membership, role, and shared projects are strongly predictive of contact patterns. Estimated reporting probabilities were low only for 0–5 min contacts. Adjusting for reporting error changed the estimate of the duration distribution, but did not change the estimates of covariate effects and had little effect on epidemic predictions. Our epidemic simulation study indicates that inclusion of network structure based on architectural and organizational structure data can improve the accuracy of epidemic forecasting models.

Keywords: *contact network, epidemic model, infectious disease, space syntax, measurement error, discordant reports, reporting error, latent variable model, social network, valued network*

1 Introduction

Influenza has a strong impact on public health, with seasonal transmission causing 3–5 million cases of severe illness and up to half a million deaths worldwide each year (WHO, 2009). Moreover, recent research has emphasized the ongoing threat of an A(H5N1) "avian" influenza pandemic with serious global health consequences (Herfst et al., 2012; Russell et al., 2012). Large-scale epidemic simulation models have been developed to predict the spread of newly evolved virus strains, as well as compare different intervention strategies: for example, the three models compared in Halloran et al. (2008). These models represent face-to-face human contacts through which influenza can be transmitted, but make assumptions about contact patterns rather than estimating contact network structures from data. For example, they assume random mixing within mixing groups known to be key to influenza transmission: households, classrooms and schools, workgroups, and workplaces. Several studies have estimated household and school contact networks with the aim of improving model specification for epidemic forecasting models, since these areas are known to be important for disease spread. Potter et al. (2011) and Potter & Hens (2013) estimate household network structures by analyzing data from contact diaries administered in a large-scale multicountry survey of contact patterns (Mossong et al., 2008). Stehlé et al. (2011) describe a face-to-face contact network in a French primary school using proximity sensor data and comment on implications for epidemic interventions. For example, the extent of class homophily in their network suggests that class-based interventions could control disease spread more efficiently than school closures. The ongoing Social Mixing And Respiratory Transmission in Schools (SMART) study has recently collected data from students in several Pittsburgh K-12 schools in order to better understand influenza transmission routes within schools (SMART Research Team, 2013). Contact data were collected from proximity sensors, contact diaries, and video-recordings of classrooms and will be used to inform epidemic models and intervention strategies employed by the Centers for Disease Control (CDC). Salathé et al. (2010) analyze wireless sensor data to describe the contact network in an American high school and demonstrate through simulation studies that using network data to inform interventions can reduce the disease burden. Potter et al. (2012) estimate network structures in a high school using data from contact diaries as well as friendship network data. Additional studies have used wireless sensors to measure contact patterns and describe network patterns in hospitals (e.g. Vanhems et al. (2013); Isella et al. (2011); Hornbeck et al. (2012)), as well as at professional conferences (Stehlé et al. (2011); Cattuto et al. (2010)), in order to better understand the impact of network structure on disease spread in these settings.

Workplaces are a potential key area of transmission which has received less attention than households and schools. Workplace-based interventions have the potential to reduce influenza attack rates, as found by Kelso et al. (2009). Furthermore, interventions exploiting the social structure within workplaces (i.e. regular workgroups), might be substantially more cost-effective, as suggested by modeling studies, than those aimed to the entire organization (Ferguson et al., 2006). A better understanding of workplace network structure will help us develop more effective and efficient interventions.

Simulation models used for epidemic prediction rely on detailed demographic and transportation data and vary somewhat in their conceptualization and construction. Some assume random mixing within mixing groups (homes, schools, classrooms, workplaces, etc.), such as Chao et al. (2010), Ferguson et al. (2006) and Milne et al. (2008). Others create activity schedules based on activity surveys, map these schedules to locations, and assume random mixing within the location (a building or a room within the building), such as Stroud et al. (2007), Smieszek et al. (2011), and Iozzi et al. (2010). While different data sources are used, none of these models use contact data to estimate or validate the workplace contact network structure. Model specification could be improved by collecting workplace contact data and using it to estimate workplace contact networks. The network model would ideally include structures relevant to the disease transmission process, but for the sake of parsimony, omit those which have no impact on epidemic dynamics. We contribute to this area by developing a network model for contacts in workplaces relevant to infectious disease transmission.

We use architectural distances measured between workstations, as well as demographic and social/organizational variables, to model contacts between members of a research institute. Several papers have explored the relationship between workplace contact and distance between desks, in the context of studying communication patterns. In their seminal study of seven R&D labs, Allen & Fustfeld (1975) identified communication patterns as a function of distance. The closer engineers were located, the higher the probability for communication was. Beyond a distance of 25 to 30 meters between workstations of a pair of engineers, their probability of communicating at least once a week decreased rapidly. These findings were confirmed in more recent studies, where (with one exception) daily face-to-face interaction in eight different knowledge-based organizations seemed to take place at a distance of 15 to 22 meters, depending on the size of the organization, its spatial configuration and office typology (Sailer & Penn, 2009; Sailer, 2010). Again, longer distances resulted in lower contact frequencies (weekly or monthly) on average. Most recently, it was argued that detailed architectural distance measures between desks of co-workers provide an important rationale for tie formation in intra-organizational interaction networks. Two actors are more likely to interact with each other when they are closely co-located, even if controlling for structural effects within networks (like transitivity and reciprocity) and organizational effects (perceived usefulness of alter and team affiliation) (Sailer & McCulloh, 2012). Our study also uses architectural distance measures to predict contact, but has a somewhat different aim than these. We focus on durations of face-to-face interactions in order to predict epidemic spread; while the others focused on communication patterns pertaining to workplace productivity.

This paper contributes two statistical innovations to the area of social network analysis. The first is in developing models for social networks with valued edges. A social network may be depicted by representing social actors as *nodes* and the social connections between them by *edges* or *ties*. We can represent a network mathematically by the *sociomatrix* $\mathbf{Y}$, defined by $Y_{ij} = 1$ if there is a tie from person i to person j and $Y_{ij} = 0$ otherwise. We refer to such a network (with 0/1 edges), as *binary*. If each edge has a value (for example, the duration of contact between i and j), we refer to the network as *valued*. A commonly used

class of network models are exponential family random graph models (ERGMs), a flexible class of model originally developed for binary networks (Strauss & Ikeda, 1990). These models allow the researcher to include effects such as homophily (the tendency for actors to associate with similar actors), and transitivity (the increased likelihood for a tie between two actors who both have a tie to a common third person). Previous research has incorporated ERGMs into latent variable models for disease transmission, in which spatial and individual data were observed but the actual contact network is unobserved (e.g. Groendyke et al. (2012)). Methods have been developed for parameter estimation and simulation of binary networks from ERGMs (Hunter & Handcock, 2006; Handcock et al., 2003). Krivitsky (2012) extends ERGMs to the case of valued networks, discusses the challenges of model specification and parameter estimation that arise, and applies these techniques to two real social networks. Hoff (2005) creates a model applicable to valued network data by including a bilinear effect and fixed and random effects in a generalized linear model and proposes a procedure for Bayesian inference. We explored fitting special cases of the class of models proposed by Krivitsky (2012) to our data and concluded that the most appropriate model for our data required categorizing durations, so creating an ordinal network. We then use a proportional odds model to model the network of duration categories. This model for ordinal network data may be applied to networks in a variety of settings.

The second statistical advancement in our work is the incorporation of reporting error into our network model. Recent simulation studies have shown that error in reports of network edges can have a substantial impact on estimation of network parameters (Žnidaršič et al., 2012; Wang et al., 2012; Almquist, 2012). Network researchers frequently note inconsistencies in edge reports, but it is fairly common to discard the information in the inconsistency patterns either by assuming that an edge exists if at least one of the two people involved reported it (e.g. Potter et al. (2012)) or restricting analysis to mutually reported edges (e.g. Goodreau et al. (2009)). When such an approach is taken for inherently symmetric networks, it may result in an underestimate of uncertainty. For inherently asymmetric networks, it simply results in a loss of information which could be analyzed explicitly. Some studies have estimated rates of inconsistent reports (e.g. Smieszek et al. (2012); Adams & Moody (2007); Helleringer et al. (2011)) but do not incorporate these estimates in a statistical framework in which they also estimate network effects. Such a framework is proposed by Butts (2003), who proposes a hierarchical Bayesian model to jointly estimate posterior distributions of network parameters and probabilities of false negative and false positive tie reports for binary networks. We instead use a likelihood-based approach, so do not impose additional assumptions implemented by the prior distributions in the Bayesian model. We also extend our model to ordinal rather than binary networks. We analyze the same data that Smieszek et al. (2012) used to estimate reporting probabilities, but we extend their model to jointly estimate reporting probabilities and network effects. We validate their findings and explore the impact of adjusting for reporting errors on network estimates and epidemic predictions.

This paper is organized as follows. In Section 2.1, we describe the contact data collected from members of a Swiss research institute, and in Section 2.2 we describe the construction of architectural distance measures computed between desks of

these members. In Section 3.1, we describe the class of network models known as ERGMs, which we employ and expand upon here. In Section 3.2, we describe our latent variable model for the binary network of contacts, which jointly estimates ERGM network effects and duration-specific reporting probabilities. In Section 3.3, we expand this model to estimate network effects for the network of categorized durations using a proportional odds model, while jointly estimating duration-specific reporting probabilities. In Section 3.4, we describe our epidemic simulation study, which explores whether our model captures the network structures important for disease transmission and compares its predictions to predictions based on random mixing. We report our estimates for the binary network model in Section 4.1 and those for the ordinal network model in Section 4.2. We report results for our epidemic simulation study in Section 4.3. In Section 5, we discuss our findings and make recommendations for future work.

2 Data

2.1 Contact survey

Longitudinal contact data and demographic information of the employees of three research groups at a Swiss university were collected using a questionnaire and contact diary approach (Smieszek et al., 2012). The three research groups belonged to one institute, and 66 individuals worked for one or more of the three groups. At least four individuals were absent from work during the entire period of data collection. Fifty individuals completed and returned both questionnaire and diary resulting in a participation rate of at least 80.6%. Contact data were collected during five consecutive workdays between Monday, May 17, 2010, and Friday, May 21, 2010.

Participants reported gender, age, research group membership(s), function within the research groups (professor, senior scientist, PhD student, administrative staff), the days on which they were usually in the office, and with whom they shared their office. Participants also reported all potentially contagious contacts they had with other participants of this study. A potentially contagious contact was defined as either conversation held at <2 m distance and with more than ten words spoken, or any sort of physical contact with another individual. For each contact, participants reported the name of the counterpart, the total contact duration during the entire day (in minutes), and whether the contact was conversational, physical, or both. Contacts were reported separately for each of the five study days. All participants were asked to complete their diaries independently and not to communicate with the other participants about the contents. For each day analyzed, we omitted from our analysis participants for whom no contacts were reported on that date, assuming this to be an indication of their absence from work.

The contact data, as well as the architectural distance data (described in the next section) are publicly available in the online supplementary material.

2.2 Architectural distance measures

This paper uses different ways of measuring the architectural distance between desks of co-workers, as initially introduced by Sailer & McCulloh (2012). In order
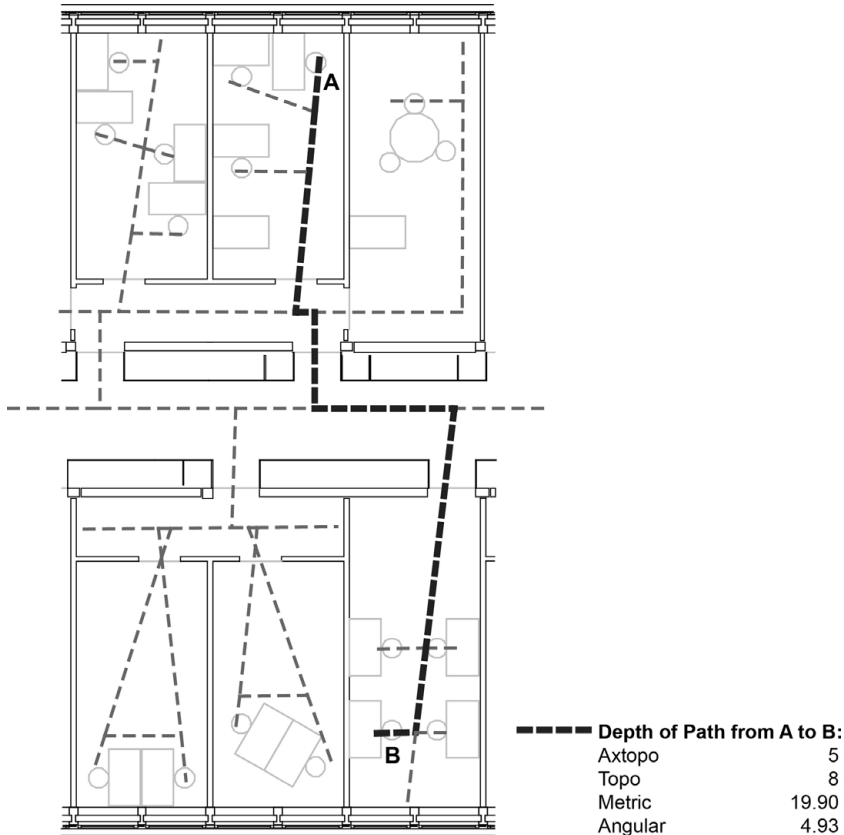
Fig. 1. An example office layout with axial topology and four architectural distance measures computed from A to B.

to represent distances, a map of lines following possible routes through the office building is drawn using Space Syntax methodologies (Hillier & Hanson, 1984; Hillier, 1996). This line map consists of all longest straight lines covering all relevant parts of the office, reaching all individual workstations and minimizing the number of lines and elements needed to go from one space to another (see Figure 1). The different floors of the office are linked through the staircases, again with lines representing the potential movement flow of people up and down the flights of the stair.

With this representation of space as a network of lines, shortest paths can be constructed from one desk to another desk. Based on Hillier & Iida (2005) and calculated using the software SEGMEN (Iida, 2009), four different distance measures can be derived from this map for the distance between any desk A and B:

1. The "axtopo distance" between two desks is the number of axes (i.e. full lines) passed on the way from one desk to another. By convention the root line is not counted.
2. The "topo distance" between two desks is the number of segments passed on the way from one desk to another. This is based on each full line broken down into separate segments at each intersection of two lines. Again, the root segment is not counted.

3. The "metric distance" between two desks is the total length of the route from one desk to another in walking meters. The distance is calculated from the center of each segment by convention.

4. The "angular distance" between two desks is the sum of changes of direction occurring on the route from one desk to another. The model assigns a 90 degree angle a weight of 1. Thus, a route with three 90 degree turns would have angular distance 3.

Examples of these four distance metrics computed on an office layout are shown in Figure 1. Previous research has shown that these distance measures capture different experiences and notions of distance (perceived distance versus actual distance) and model an existing environment more accurately than the most commonly used Euclidean distances. For more details on this discussion, see Sailer & McCulloh (2012).

3 Methodology

3.1 Exponential family random graph models

We use exponential random graph models, following the example of Sailer & McCulloh (2012) to estimate the effects of individual attributes and architectural distances on contact patterns. We also expand on ideas set forth in Smieszek et al. (2012), who use these data to estimate the probability of reporting an existing contact. We develop a model to estimate ERGM parameters for the true network of contacts, statistically correcting for measurement error and inconsistencies in reporting. We do this by expressing the likelihood of the network of reported contacts in terms of the network of true contacts and the probability of reporting a contact. Parameter estimates are then obtained through maximum likelihood estimation.

We first define some network terminology and notation. We can depict the network by representing the social actors by nodes and the contacts between them with edges or ties. A pair of actors is called a dyad. We represent the observed contact network by a sociomatrix, C, a square matrix with as many rows as there are people in the network, where $C_{ij} = 1$ if i reports contact to j and 0 otherwise. We represent the sociomatrix of true contacts by Y, so $Y_{ij} = 1$ if i and j actually made contact, regardless of whether that contact was reported, and $Y_{ij} = 0$ if no contact was made. Because of inconsistencies in reporting, C is an asymmetric matrix. However, the sociomatrix of true contacts, Y, is symmetric because contact by our definition is symmetric. Let the four duration categories 0–5, 6–15, 16–60, or 60–480 min be denoted by d_k, for $k \in \{1, 2, 3, 4\}$, and let p_k denote the probability of reporting an existing contact of duration d_k. We assume the reporting probability is independent of the contact probability for any two actors. We assume, as in Smieszek et al. (2012), that the events that different respondents report existing contacts are independent, and that no contacts were fabricated. We also assume that the reporting probability does not depend on covariates which may predict contact patterns. This assumption is comparable to a *missing completely at random* (MCAR) assumption if reporting errors are viewed as missing data. We explored modeling such a dependency (discussed below) but found this not to improve our model.

We would like to jointly estimate **p** and Y. Since Y is unobserved, we will use a latent variable model for estimation. We will express the likelihood of our observed data C in terms of Y and p and compute the maximum likelihood estimate.

We use ERGMs to model the true contact network Y. An ERGM takes the form

$$P(Y = y|\theta) = \frac{e^{\theta^T \mathbf{g(y)}}}{\kappa(\theta, \mathscr{Y})}$$

Above, $\mathscr{Y}$ denotes the set of all networks of this size, $\kappa(\theta, \mathscr{Y})$ is a normalizing constant ensuring that the probability distribution sums to one, θ is a vector of parameters, and $\mathbf{g(y)}$ is a vector of network statistics capturing network structures we want to estimate. For example, $\mathbf{g(y)}$ may include an edges term for a density effect, the number of contacts between members of the same sex for a mixing effect, or a triangles term to capture transitivity (the increased likelihood of two people who have mutual contacts to make contact). The parameters θ are estimated with the maximum likelihood estimate (MLE). In general, the normalizing constant does not have an analytic form, so the MLE is approximated with an MCMC procedure (Snijders, 2002).

The choice of statistics in $\mathbf{g(y)}$ specifies the model. Some ERGMs are *dyad-independent*, which means that the event of contact occurring on one dyad is independent of contact patterns on other dyads. In dyad-independent models, contact behavior is characterized by individual-level and dyadic attributes, and the MLE may be estimated with logistic regression rather than MCMC. In *dyad-dependent* models, $\mathbf{g(y)}$ includes dependency terms, such as the number of triangles.

We included the following statistics in our ERGM:

1. The number of edges (a density effect).
2. Two terms to estimate sociality effects for each research group: the number of contacts made by members of research group 1 and the count for group 2. Group 3 is used as the reference group, so these terms estimate how much more social members of groups 1 and 2 are than members of group 3.
3. The number of contacts made between members of the same research group, estimating a preference to contact others in the same group. This effect is distinct from the previous one, because while some groups may be more social than others, their contacts may occur in different ratios of between versus within-group contacts.
4. The total distance between members making contact. We fit four separate ERGMS with four separate distance metrics.
5. Sociality effects by gender: the number of contacts made by females.
6. Gender homophily: the number of same-gender contacts.
7. Class homophily: the number of contacts between members of the same function (graduate students, postdocs, or administrative staff). No professors participated in the survey.
8. The total number of contacts between people on the same floor. People may be more inclined to contact others on the same floor.
9. Shared-projects homophily: the total number of contacts between people who work on the same projects together (weighted as 1 or 2, depending on whether they are mutually reported).

The ERGMs we selected are dyad-independent, in which case the likelihood of the actual network is equivalent to logistic regression with dyads as the dependent variable. The assumption of dyad-independence is a strong one, since additional clustering may be present in the network which is not explained by the various mixing effects included in our model. Adjusting for reporting errors with a dyad-dependent ERGM would be extremely complex mathematically and computationally, so we chose to begin with the simpler dyad-independent models. We express the probability distribution of one dyad as:

$$\text{logit}(P(Y_{ij} = 1)) = \beta_0 + \beta_1(1_{[\text{i in group 1}]} + 1_{[\text{j in group 1}]})$$

$$+\beta_2\,(1_{[\text{i in group 2}]} + 1_{[\text{j in group 2}]}) + \beta_3 1_{[\text{i, j in same group}]}$$

$$+\beta_4\,\text{distance}(i, j) + \beta_5(1_{[\text{i female}]} + 1_{[\text{j female}]}) + \beta_6 1_{[\text{i,j same sex}]}$$

$$\beta_7 1_{[\text{i,j same function}]} + \beta_8 1_{[\text{i,j on same floor}]}$$

$$+\beta_9\,(1_{[\text{i reports shared projects with j}]} + 1_{[\text{j reports shared projects with i}]})$$

When contacts are symmetric and dyads are independent, we obtain:

$$P(Y = y) = \prod_{i=1}^{n} \prod_{j=i+1}^{n} P(Y_{ij} = y_{ij})$$

3.2 Likelihood of reported contacts, duration-specific reporting probabilities

Next, we express the likelihood of the reported contacts, C. We expect reporting probability to vary with duration of contacts; Smieszek et al. (2012) found that shorter contacts were more likely to be forgotten. In this section, we jointly model the reported contacts and durations, and we estimate separate reporting probabilities for each duration category: 0–5, 6–15, 16–60, or 60–480 min. We denote the four duration categories by d_k, for $k \in \{1, 2, 3, 4\}$. As in Smieszek et al. (2012), when two participants reported different durations for the same contact, we assume that the longer duration is the correct report. We also assume that duration of contact does not depend on individual or dyadic attributes, and we again assume independence in contact (and durations) between dyads, conditional on the effects in our model. Let γ_k denote the probability of an existing contact having duration d_k, so $\gamma_k = P(D_{ij} = d_k | Y_{ij} = 1)$. Let D denote the matrix of contact durations, after removing inconsistencies in duration reports, so D is a symmetric matrix. By applying our assumptions (including dyad independence), rules of conditional probability, and the Law of Total Probability, we find that the joint likelihood of D_{ij} and C_{ij} is

$$P(C_{ij} = 1, C_{ji} = 1, D_{ij} = d_k) = \gamma_k P(Y_{ij} = 1)p_k^2$$

$$P(C_{ij} = 0, C_{ji} = 1, D_{ij} = d_k) = \gamma_k P(Y_{ij} = 1)p_k(1 - p_k)$$

$$P(C_{ij} = 0, C_{ji} = 0) = P(Y_{ij} = 0) + \sum_{k=1}^{4} \gamma_k P(Y_{ij} = 1)(1 - p_k)^2$$

Then, the probability of observing the reported contact network is found by using the above equations to express the probabilities in the following formula:

$$P(C = c, D = d) = \prod_{i=1}^{n} \prod_{j=i+1}^{n} P(C_{ij} = c_{ij}, C_{ji} = c_{ji}, D_{ij} = d_k)$$

We maximized the log likelihood using R software with the `trust` function in R (Geyer, 2009; R Development Core Team, 2011). This optimization method requires gradient and Hessian functions of the log likelihood as input values, and we approximated these with the `grad` and `hessian` functions in the `numDeriv` package in R (Gilbert, 2011). The optimization routine returns the parameter vector maximizing the log likelihood as well as the value of the Hessian at the MLE. We computed standard errors by inverting the negative of the Hessian (the observed Fisher information matrix).

3.3 *Likelihood of contact durations with duration-specific reporting probabilities*

The duration of each contact may depend on individual and dyadic attributes. For example, short contacts may occur more frequently between those who share an office. The idea is that if one travels an extra distance to contact another person, they are likely to make a longer contact. In the previous section, we assumed that durations were uniformly distributed; in this section we refine that model to estimate duration from known attributes.

In order to do this, we need to create a model for the probability distribution of the duration matrix and derive the expression for its likelihood. Then, we will express the joint likelihood of the true duration matrix and the reported duration matrix as in previous sections, and maximize the log likelihood function with `trust`. Our reported durations have a large number of zeroes and are overdispersed. The mean of the nonzero duration reports is 26 min, and variance 987. We could use a generalized linear model to estimate the mean of the duration distribution as a function of covariate values (McCullagh & Nelder, 1989). For this approach, we considered a negative binomial distribution and a zero-inflated negative binomial distribution (fits shown in Figure 2). Our actual distribution of contact durations has spikes at 30, 60, and 90 min, either because people tended to round their durations to these values or because based on common meeting lengths, these values are actually more frequent. The parametric forms we considered did not capture this phenomenon, so we instead categorized duration in order to avoid imposing assumptions on the duration distribution.

We used two methods to estimate probabilities of a duration falling into each category as a function of covariates. The first approach was multinomial logistic regression, which has no distributional assumptions but does not take advantage of the fact that duration categories are ordered. The second method was a proportional odds model, which does exploit the ordering, but imposes an additional assumption. We found the proportional odds model to be more appropriate for this data set. The multinomial model created some estimation problems due to the large number of parameters and some cases of very small cell counts. While the multinomial model allows additional flexibility, the results did not show strong evidence that effects varied by duration category. Since the additional flexibility did not yield
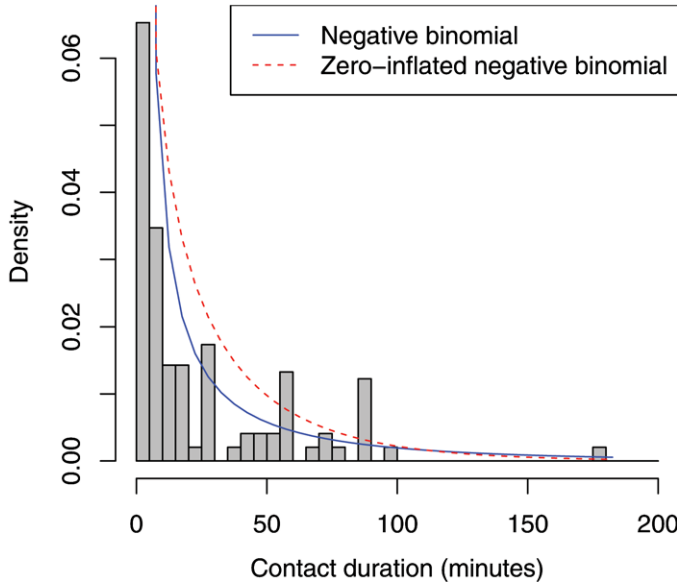
Histogram of contact durations

Fig. 2. The observed distribution of contact durations, with negative binomial and zero-inflated negative binomial fitted models. Contacts with zero duration (i.e. non-contacts) comprised 92% of all dyads and are omitted from the graph. (color online)

extra insight, we decided that the proportional odds model is preferable, so we restrict our attention here to that model. A detailed description of the multinomial model and its performance is included in the supplementary material.

The proportional odds model is defined by

$$\text{logit}(\text{P}(D_{ij} \leqslant d_k)) = \alpha_k - \boldsymbol{\beta}^T \mathbf{x}$$

While the multinomial model estimates a separate vector of covariate effects for each duration category, here a single vector of covariate effects is estimated. The indexing variable k takes values from zero to four, for each possible duration category (0, 1–5, 6–15, 16–60, >60).

This model satisfies:

$$\text{logit}(\text{P}(D_{ij} > d_k | \mathbf{x_1})) - \text{logit}(\text{P}(D_{ij} > d_k | \mathbf{x_2})) = \boldsymbol{\beta}^T (\mathbf{x_1} - \mathbf{x_2}),$$

that is the log cumulative odds ratio is a fixed linear combination of the differences in the covariate values, and this linear combination is the same for each category of the outcome variable. The probability of observing duration category k for contact between i and j is

$$P(D_{ij} = d_0 | X_{ij} = \mathbf{x}) = \frac{e^{\alpha_0 - \boldsymbol{\beta}^T \mathbf{x}}}{1 + e^{\alpha_0 - \boldsymbol{\beta}^T \mathbf{x}}}$$

$$P(D_{ij} = d_k | X_{ij} = \mathbf{x}) = \frac{e^{\alpha_k - \boldsymbol{\beta}^T \mathbf{x}}}{1 + e^{\alpha_k - \boldsymbol{\beta}^T \mathbf{x}}} - \frac{e^{\alpha_{k-1} - \boldsymbol{\beta}^T \mathbf{x}}}{1 + e^{\alpha_{k-1} - \boldsymbol{\beta}^T \mathbf{x}}}, \quad \text{for } k = 1, 2, 3$$

$$P(D_{ij} = d_4 | X_{ij} = \mathbf{x}) = 1 - \sum_{k=0}^{3} P(D_{ij} = d_k | X_{ij} = \mathbf{x})$$

By applying our assumptions, rules of conditional probability, and the Law of Total Probability, we find that the joint likelihood of D and C is

$$P(C_{ij} = 1, C_{ji} = 1, D_{ij} = d_k) = P(D_{ij} = d_k)p_k^2$$

$$P(C_{ij} = 1, C_{ji} = 0, D_{ij} = d_k) = P(D_{ij} = d_k)p_k(1 - p_k)$$

$$P(C_{ij} = 0, C_{ji} = 0, D_{ij} = 0) = P(D_{ij} = 0) + \sum_{k=1}^{4} P(D_{ij} = d_k)(1 - p_k)^2$$

Then, the probability of the observed data is

$$P(C = c, D = d) = \prod_{i=1}^{n} \prod_{j=i+1}^{n} P(C_{ij} = c_{ij}, C_{ji} = c_{ji}, D_{ij} = d_k)$$

We maximized the log likelihood to estimate $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\mathbf{p}$ using the trust function in R and computed standard errors by inverting the Fisher information matrix (Geyer, 2009).

We tested the proportionality assumption by dichotomizing the outcome variable at each of the duration cutpoints, fitting separate logistic regression models, and comparing odds ratio estimates from the different models. In testing the assumption, we did not jointly estimate reporting probabilities together with covariate effects because extending our model is not straightforward when contact duration is dichotomized. Instead, we assumed that contact occurred if either or both members reported it. These results are included in the supplementary material and indicate that the proportional odds assumption is reasonable.

We considered modeling the reporting probability as a function of covariates. For example, older people may be more likely to forget contacts than younger people. Exploratory data analysis revealed that members are less likely to forget contacts to those who work on a different floor, but did not show evidence for other covariate effects on reporting probability. Among those on the same floor, 17% of reports were consistent, while 82% of contact reports between members on different floors were consistent. We extended the proportional odds model to include a floor effect on reporting probability. We estimated the log odds ratio of reporting probability for different to same floors, and assumed the same odds ratio for different contact durations. However the estimated log odds ratio was not statistically significant (95% C.I. [−1.23, 0.49]), so we decided to omit this effect from our final model. Our power is limited because there were only eleven contact reports between members on different floors.

3.4 Simulation model of epidemics

Our epidemic simulations were based on an individual-based model of influenza spread used in previous publications (Salathé et al., 2010; Smieszek & Salathé, 2013). The model is a stochastic SEIR model with simulation time steps of half a day. We assume that infection is introduced by one randomly chosen index case at

the beginning of a simulation run and that, after the initial introduction, there are no further introductions from outside.

When a susceptible individual has contact with one or more infectious individuals, the probability to switch from the susceptible to the exposed state is $1 - (1 - \psi)^w$, where ψ is the probability of transmission per minute and w is the accumulated contact time the susceptible individual has spent with infectious individuals during the entire half-day at work. Previous work estimated the transmission probability for an influenza outbreak that occurred on a plane to be $\psi = 0.009$ min^{-1} (Moser et al., 1979). We used this as an initial input for our simulations. As influenza strains vary in their infectiousness, we performed additional simulations using fifteen different infectivity parameters, ranging from $\psi = 0.009$ min^{-1} to $\psi = 0.135$ min^{-1}. The duration of the exposed state follows a Weibull distribution with an offset of half a day; the power parameter is 2.21, the scale parameter is 1.10 (Ferguson et al., 2005). After the exposed state, each individual remains in the infectious state for exactly one time step, and then withdraws to the home, so is removed from the workplace population.

A week in our simulation model consists of 14 timesteps: five workdays with each one time step at the workplace and one time step at home as well as two weekend days with in total four time steps at home. Since workplace transmission can only occur during the five time steps at work, any exposed individual turning infectious at home, will not be able to pass on the infection to colleagues at work. Unlike previous applications of this model (Salathé et al., 2010; Smieszek & Salathé, 2013), we assume here that the initial case will become infectious during one of the five workday time steps of a simulation week, which is determined randomly for each run.

In order to assess the adequacy of our model to represent the actual network structure, we compare epidemic simulations based on simulated contact networks from our model to those based on the original data collected on the Monday of the study week. For simulations based on the original data, we assume that contact occurred if it was reported by one or both members. We also compare these two results to simulations based on a version of our model which does not adjust for inconsistent reports, and which uses the "standard assumption" which converts discordant ties to concordant ones.

We compare our network-based epidemic simulation results to a random mixing model, commonly used in epidemic predictions. To make a fair comparison, the random mixing model has the same total number of contacts (on average) as a network simulated from our model, and durations in the random mixing model are all equal to the mean duration. We also perform epidemic simulations based on subsets of our model in order to assess whether a more parsimonious model would sufficiently represent the network structure relevant to epidemic spread. Finally, we perform simulations based on contact networks generated by random reshuffling of the edges generated from one simulation of the full model. To this end, two pairs of nodes were randomly chosen, and their contact duration (including no contact operationalized as contact of zero duration) was swapped. This was repeated 100,000 times. The reshuffling model preserves the distribution of edge duration alone, so testing the effect of this single network structure.

Table 1. *Descriptive statistics of 50 members of the Swiss research institute.*

Variable	Mean(SD) or frequency table
Age	31.7 (6.6)
Sex	25 Male
	25 Female
Role	15 Postdoctoral fellows
	24 Graduate students
	6 Administrative staff
Group	24 Group 1
	19 Group 2
	11 Group 3

We generated 500 different realizations for all contact network models and computed 10,000 individual simulation runs for each realization.

4 Results

Table 1 shows descriptive statistics of our data set. There are equal numbers of men and women, and most members are graduate students or postdocs. No professors responded to the survey. One member belonged to all three groups and two members belonged to both groups one and two.

4.1 Results for models for binary contact network

Table 2 shows parameter estimates for our model for the binary contact network fit with four different distance metrics. The best-fitting model according to the AIC is the model including angular distance. Estimates for this model are interpreted as follows: the odds of a contact stemming from a member of group 2 is $e^{0.38} = 1.46$ times the odds of contact stemming from a member of group 3 (the reference group), controlling for other terms in the model. The odds of contact is increased by a factor of $e^{3.48} = 32$ if the two people belong to the same research group, controlling for other effects. For each additional unit of angular distance between two workstations, the odds of contact decreases by a factor of $e^{-0.27} = 0.76$, controlling for other effects in the model. The odds of contact is increased by a factor of $e^{0.14} = 1.15$ if it stems from a female rather than from a male, although this effect is not significant. The odds of contact increases by a factor of $e^{1.44} = 4.22$ if they share work projects. Effects in the other models are interpreted similarly. The coefficient for shared projects in the metric distance and axtopo distance models is effectively infinite, meaning that once controlling for other effects in the model, those who share projects always make contact. The angular and axtopo distance models perform similarly, because the construction of these distance metrics is similar. The metric and topo models perform similarly for the same reason.

Table 2. *Coefficient estimates for logistic regression models of binary contact network with four different distance metrics.*

Estimate	Architectural distance metric included in ERGM			
	Metric	Angular	Topo	Axtopo
Group 1	−0.34 (0.24)	0 (0.23)	−0.35 (0.24)	−0.04 (0.22)
Group 2	0.15 (0.23)	0.38 (0.23).	0.14 (0.22)	0.40 (0.23).
Group mixing	3.75 (0.55)***	3.48 (0.48)***	3.77 (0.54)***	3.44 (0.48)***
Distance	0.01 (0.02)	−0.27 (0.11)*	0.04 (0.05)	−0.23 (0.10)*
Female	0.26 (0.25)	0.14 (0.24)	0.26 (0.25)	0.14 (0.24)
Role mixing	0.68 (0.36).	0.60 (0.34).	0.70 (0.36).	0.63 (0.34).
Gender Mixing	−0.22 (0.31)	−0.23 (0.30)	−0.21 (0.31)	−0.23 (0.30)
Floor	1.15 (0.59).	-0.69 (0.81)	1.38 (0.70).	−0.93 (0.94)
Shared projects	19.02 (NA)	1.44 (0.53)**	21.46 (NA)	1.47 (0.54)**
Reporting probability				
0–5	0.48 [0.36, 0.60]	0.56 [0.41, 0.69]	0.48 [0.36, 0.60]	0.55 [0.41, 0.69]
6–15	0.96 [0.84, 0.99]	0.96 [0.84, 0.99]	0.96 [0.84, 0.99]	0.96 [0.84, 0.99]
16–60	0.93 [0.83, 0.98]	0.93 [0.83, 0.98]	0.93 [0.83, 0.98]	0.93 [0.83, 0.98]
61–480	1.00 [0, 1.00]	1.00 [0, 1.00]	1.00 [0, 1.00]	1.00 [0, 1.00]
Duration category				
0–5	0.49 [0.39, 0.59]	0.45 [0.35, 0.55]	0.49 [0.39, 0.59]	0.46 [0.35, 0.56]
6–15	0.18 [0.12, 0.25]	0.20 [0.12, 0.27]	0.18 [0.12, 0.25]	0.20 [0.12, 0.27]
16–60	0.24 [0.16, 0.31]	0.25 [0.17, 0.33]	0.24 [0.16, 0.31]	0.25 [0.17, 0.33]
61-480	0.09 [0.04, 0.14]	0.09 [0.04, 0.15]	0.09 [0.04, 0.14]	0.09 [0.04, 0.15]
AIC	796	789	795	790

Significance levels: *** $= p < 0.001$; ** $= p < 0.01$; * $= p < 0.05$; "." $= p < 0.10$.

The four models estimate similar reporting probabilities and duration categories. The reporting probability estimates were nearly identical to those obtained by Smieszek et al. (2012), validating our method. Both sets of estimates are compared in the supplementary material.

Table 3 compares the estimated duration distribution from the model with angular distance to the duration distribution of reported contacts. Our model necessarily estimates higher numbers of contacts than are observed, since some non-contacts are attributed to reporting errors but no contacts are considered erroneously reported. Since a much lower reporting probability (0.56) is estimated for 0–5 min contacts than for longer duration contacts (0.93–1.00), our model estimates a higher proportion of 0–5 min contacts than what is observed.

4.2 Results for proportional odds models for network of contact durations

The proportional odds model with angular distance was also best according to the AIC, so we present only that one here, although the others are included in the supplementary material. Table 4 compares estimates from our model to those from a model which does not adjust for reporting errors (the "standard model") but instead assumes that contact occurred between two people if and only if at

Table 3. *Duration distribution estimates from our model compared to the observed duration distribution of reported contacts.*

Duration category	Our estimate of true distribution	Distribution of reported contacts
0–5	0.45 [0.35, 0.55]	0.41 [0.25, 0.56]
6–15	0.20 [0.12, 0.27]	0.22 [0.00, 0.47]
16–60	0.25 [0.17, 0.33]	0.28 [0.06, 0.49]
> 60	0.09 [0.04, 0.15]	0.10 [0.00, 0.49]

Table 4. *Coefficient estimates for proportional odds model, compared to estimates using standard assumption, angular distance measure.*

	Our model	Standard model
Intercepts	Est. (SE)	Est. (SE)
Duration = 0	2.65 (1.04)*	2.81 (0.72)***
Duration ⩽ 5	3.87 (1.04)***	3.76 (0.72)***
Duration ⩽ 15	4.58 (1.04)***	4.47 (0.72)***
Duration ⩽ 60	6.22 (1.08)***	6.10 (0.74)***
Coefficients		
Group 1	−0.18 (0.20)	−0.16 (0.14)
Group 2	0.11 (0.19)	0.12 (0.13)
Group mixing	3.42 (0.45)***	3.34 (0.31)***
Distance	−0.22 (0.08)**	−0.22 (0.06)***
Female	0.31 (0.21)	0.29 (0.14)*
Role mixing	0.60 (0.29)*	0.58 (0.20)**
Gender mixing	−0.18 (0.26)	−0.17 (0.18)
Floor	−0.09 (0.68)	−0.13 (0.47)
Shared projects	1.06 (0.28)***	1.07 (0.19)***

Significance levels: $*** = p < 0.001$; $** = p < 0.01$; $* = p < 0.05$; ".” $= p < 0.10$.

least one of the two reported it. Once again, duration distributions differ slightly (though not significantly), but effect estimates are nearly identical. The standard errors for our model are larger than those for the standard model, because they incorporate the additional uncertainty contributed by reporting errors. Coefficients for this model are interpreted as follows: The odds of duration being greater than a certain category increases by a factor of $e^{3.42} = 30.6$ when two members are in the same group. The odds of duration being greater than a specified category decreases by a factor of $e^{-0.22} = 0.80$ with each additional unit of angular distance between their workstations.

Table 5 shows estimates for subsets of the full model. We performed epidemic simulations over these more parsimonious models to compare them to the full model and assess whether one of them adequately captured all network structure relevant for epidemic predictions.

Table 5. *Coefficient estimates for proportional odds model and subsets of model, angular distance measure.*

	Model 1	Model 2	Model 3	Model 4	Full Model
0	−0.48 (0.71)	3.93 (0.39)***	3.92 (0.39)***	2.67 (1.04)*	2.65 (1.04)*
1–5	0.45 (0.70)	4.95 (0.40)***	4.89 (0.40)***	3.88 (1.05)***	3.87 (1.04)***
6–15	1.03 (0.69)	5.56 (0.42)***	5.45 (0.41)***	4.58 (1.05)***	4.58 (1.04)***
16–60	2.49 (0.72)**	7.10 (0.49)***	6.89 (0.48)***	6.23 (1.08)***	6.22 (1.08)***
Coefficients					
Group 1					−0.18 (0.20)
Group 2					0.11 (0.19)
Group mixing		3.65 (0.41)***	3.87 (0.41)***	3.37 (0.44)***	3.42 (0.45)***
Shared projects		1.31 (0.27)***		0.98 (0.27)***	1.06 (0.28)***
Distance	−0.36 (0.07)***			−0.22 (0.08)**	−0.22 (0.08)**
Floor	0.69 (0.47)			0.14 (0.62)	−0.09 (0.68)
Female				0.17 (0.18)	0.31 (0.21)
Role mixing				0.61 (0.29)*	0.60 (0.29)*
Gender mixing				−0.20 (0.26)	−0.18 (0.26)
AIC	894	805	803	771	773

Table 6. *Kendall's τ computed for the rank order of individuals (according to the expected number of cases generated) between various models.*

	ψ (1/min)	Full model	Standard model	Model 1	Model 2	Model 3	Model 4
Original	0.009	0.33	0.36	0.21	0.14	0.04	0.37
	0.045	0.41	0.40	0.28	0.19	0.17	0.40
	0.090	0.44	0.43	0.27	0.19	0.26	0.43
	0.135	0.48	0.46	0.29	0.19	0.22	0.47
Full model	0.009	n/a	0.95	0.45	0.39	0.42	0.78
	0.045	n/a	0.95	0.47	0.38	0.43	0.78
	0.090	n/a	0.97	0.47	0.36	0.55	0.80
	0.135	n/a	0.97	0.50	0.38	0.52	0.82

4.3 Epidemiological properties of the contact networks and results of the epidemic simulations

4.3.1 Correlation of individual epidemiological importance

We analyzed to what extent the individual epidemiological importance of the members of the workplace population is correlated among the different contact network models. Here, epidemiological importance is operationalized as the mean expected number of cases generated by a specific individual, given that all of its workplace contact partners are fully susceptible. We calculated Kendall's τ as a robust, non-parametric measure of correlation. If we compare the rank order of the individuals (according to the expected number of cases generated) for two different types of contact networks, then $\tau = 1$ means that the rank order is identical. Contrary, $\tau = 0$ indicates that the ranks of individuals are unrelated between two networks. All other values of are as easy to interpret since the odds ratio of concordant to discordant pairs of observations is given by $\frac{1+\tau}{1-\tau}$ (Noether, 1981). Table 6 shows Kendall's τ for all pairs between (i) the original data and all other contact networks as well as (ii) the full network model and all other contact networks.

The values are interpreted as follows: when comparing the individual epidemiological importance for the full model and the original data, the odds of concordant rank orders for any pair of observations was approximately twice the odds of a discordant one (since $\frac{1+0.33}{1-0.33} = 1.99$). Comparing model 1 and the original data, the odds of a concordant pair was approximately 1.5 times that of a discordant pair. A comparison of the full and the standard model resulted in an odds ratio of 39, indicating that the correction for underreporting had very little effect on the individuals' ranking.

4.3.2 Simulated epidemics

Figures 3–5 show the mean final size for epidemic simulations based on various models. We estimated 95% confidence intervals for the mean final outbreak size using a nonparametric bootstrap (1,000 bootstrap resamples were drawn), but these were so narrow that we omit them from the graphs.
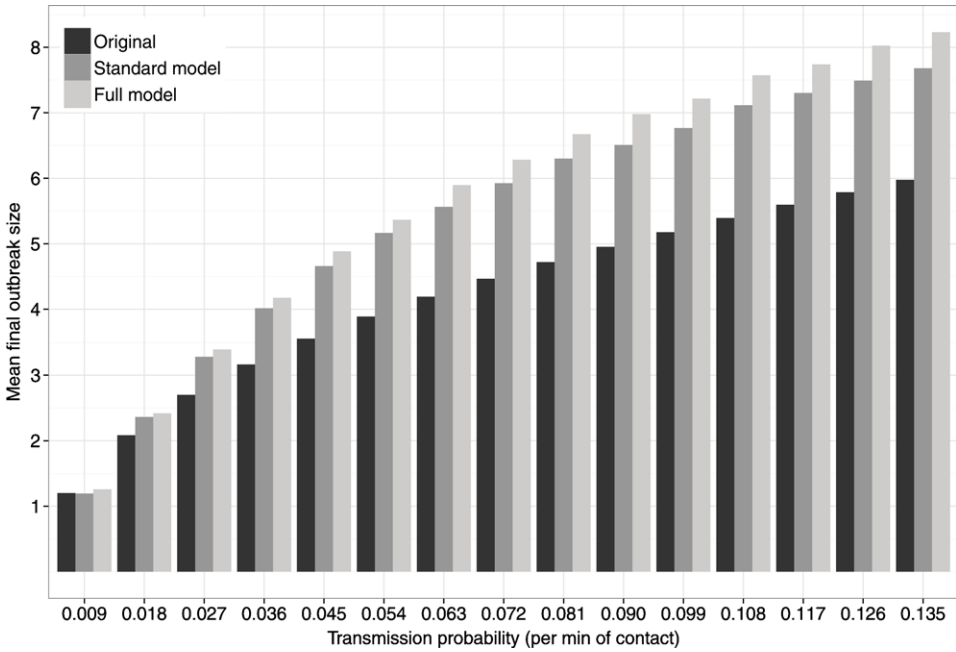
Fig. 3. Mean final size (minus index case) by transmission probability per minute of contact and for different contact networks, based on simulations. "Original" refers to the empirically measured network with reporting inconsistencies resolved; 'Full model' and "Standard model" refer to the corresponding models parameterized in Table 6.

Figure 3 compares predictions based on our model to those based on the original data. We make this comparison as a way of assessing model fit: if we have sufficiently modeled network structure relevant to disease spread, then we expect similar predictions from the two models. However, a shortcoming of predictions based on the empirical data is that they do not adjust for inconsistent reports, and thus, underestimate overall density. Therefore, we expect larger outbreaks from our model even if it does capture all relevant network structures. To adjust for this, we also include our "standard model" (the proportional odds model not adjusting for reporting errors). If our model captures all relevant network structure important for epidemic forecasting, the standard model should produce similar predictions to the original data. Figure 3 shows that the adjustment for reporting error makes only a small difference in epidemic predictions, and that we have failed to capture some of the network structure important for epidemic forecasts. However, our model's predictions are close to those based on the original data, suggesting that it does a reasonable job.

Figure 4 shows the expected final size for epidemic simulations based on the original (empirical) contact data, the full model, networks created by shuffling the edges of the full model, and a random mixing scenario. The comparison to the shuffled edges model shows the effect of the distribution of edge duration alone, since that is the only network effect included in the shuffled edges model. The figure shows that a large part of the difference between final size estimates based on random mixing and those based on the original data is due to the network effect of heterogeneity in edge duration. Additional effects included in our model account
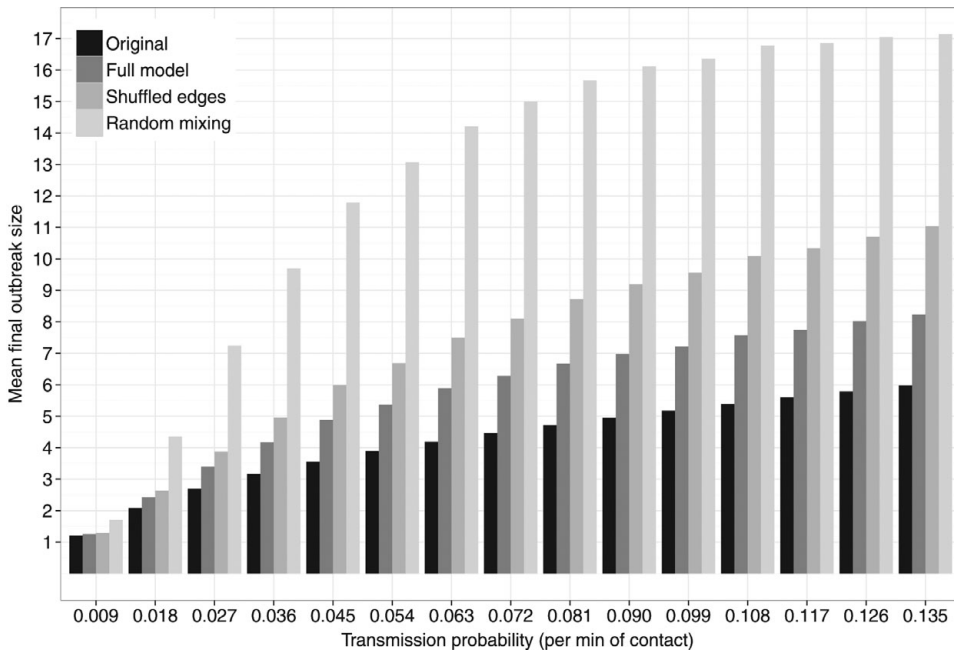
Fig. 4. Mean final size (minus index case) by transmission probability per minute of contact and for different contact networks, based on simulations. "Original" refers to the empirically measured network with reporting inconsistencies resolved; "Full model" refers to the model in Table 6; "Shuffled edges" are network models with the same density and duration distribution as the full model, but with randomly allocated edges; "Random mixing" is a random mixing scenario.

for part of the additional difference, and the remaining small difference between our full model's predictions and those based on the original data remain unexplained. All three of the network models predict smaller epidemics than a random mixing scenario, in line with what previous researchers have found (e.g. Eames (2008); Potter et al. (2012); Smieszek et al. (2009); Szendroi & Csányi (2004); Duerr et al. (2007)). The clustering and repetition found in realistic contact networks tends to slow disease spread by reducing the number of susceptible persons that each infected individual comes into contact with.

Figure 5 provides a model comparison between the full model and models 1–4 (see Table 5), which are subsets of the full model. While the results of all five models are very close, it is clear that the disease dynamics are closest to that of outbreaks on the original data for the full model and model 4, which differs from the full model only in its exclusion of group sociality effects. Model 1, which only includes architectural information, performs worse than models 2 and 3, which only include organizational information.

5 Discussion

In this paper, we have developed social network models for face-to-face contacts relevant to infectious disease transmission in a Swiss research institute. Our models use architectural distances between workstations, demographic variables,
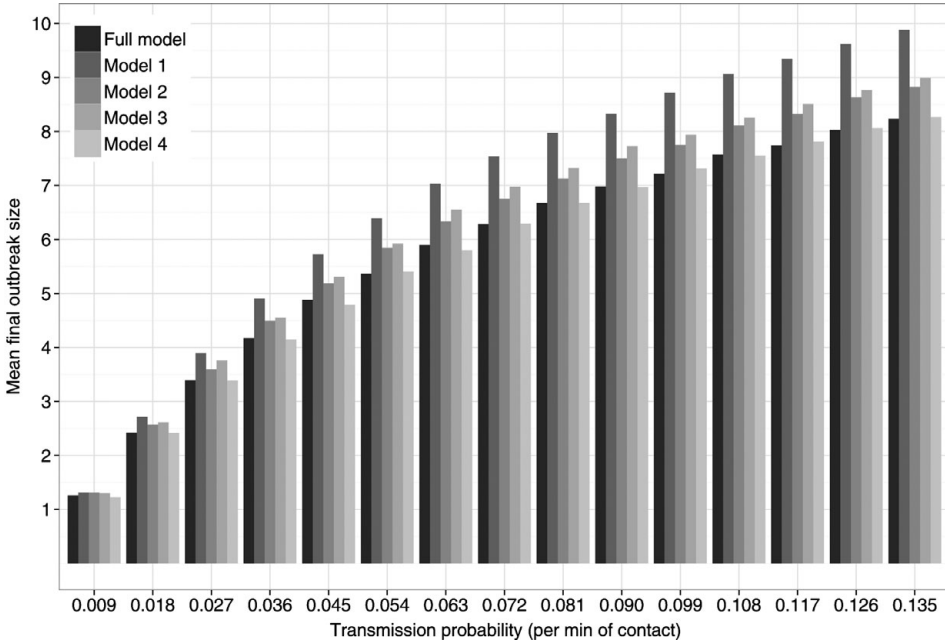
Fig. 5. Mean final size (minus index case) by transmission probability per minute of contact for the full model and for various subsets of the full model, as defined in Table 5, based on simulations.

and organizational structure to predict contact patterns. We found workgroup membership, collaboration on projects, mixing by employee role, and distance between workstations to be highly predictive of contact. We found models with angular and axtopo distance measures to have higher predictive power those with metric or topo measures. We developed latent variable models to jointly estimate duration-specific reporting probabilities and the network of contacts. We found very high reporting probabilities for contacts with duration greater than five minutes, but only 53% probability of reporting a 0–5 min contact, consistent with findings in Smieszek et al. (2012). We also extended our model to estimate the network of contact durations rather than the binary contact network. Effect estimates were similar between the two models, so (for example) while collaborating on projects increases the odds of contact, it also increases the odds of longer duration contacts.

By adjusting for measurement error and comparing to models that do not make this adjustment, we have contributed to the area of social network analysis by making a statistical improvement which can be applied to network analysis in many different settings. Our findings have several implications for scientists. First, our duration distribution estimate differs from that obtained by a model that does not adjust for reporting errors, since reporting probabilities vary by duration category. Our duration distribution estimate is more accurate. However, since reporting probabilities were very high for most durations of contacts, the difference is small, and in fact is not statistically significant. Second, in both the binary network model and the duration network model, we found effect estimates to be only slightly (and not significantly) different from those by a model which does not adjust for

reporting errors. This finding is appealing since making this adjustment requires a more complex model, programming time, and computation time. However, we believe that this finding is partly due to exceptionally accurate reporting and holds only under certain conditions. Other studies have found substantially less accurate reporting (Read et al., 2012; Smieszek et al., 2014). In our model, we assumed the reporting probability does not depend on covariate values. We believe that if the true reporting probability depended on covariate values, then adjusting for reporting errors would change network effect estimates. We believe that such a condition is quite plausible. For example, age may be related to reporting probability, as older subjects may be more likely to forget contacts. Type of job may be as well, with those in busier and more stressful positions being more likely to forget contacts. We explored such an extension, and our data suggest that subjects may be less likely to forget contacts occurring on different floors, but the effect was not statistically significant. We found no evidence for a relationship between any other covariate and reporting probability. However, our sample size is fairly small, limiting our power to detect such a relationship. It is quite plausible that such a relationship could be detected in a larger, richer data set. In summary, adjusting for reporting errors had only a small impact on estimates of duration distribution and network effects in this setting, but may have an impact in different settings. We recommend further exploration into this area, including assessment of dependence between the reporting probability and covariates in the model. In networks with lower reporting probabilities, adjustment for measurement error will be more important. A final additional statistical improvement contributed by adjusting for reporting errors is that standard errors for network effects now include uncertainty contributed by reporting errors, so are more accurate than those in the standard model. They are only slightly larger than those from a standard model, so the standard assumption does not do much harm. This is because reporting probabilities are fairly high.

Our epidemic simulation study showed that all network models considered produced epidemic forecasts closer to those based on the original data than a random mixing assumption, which tends to overestimate final size. Predicted final sizes based on our full model were similar to those based on the original data but were not identical. Part of this is due to the fact that our model adjusts for reporting errors, changing the estimate of the duration distribution. The remaining difference suggests that our model omits some network structures relevant to epidemic forecasting. We recommend further research in this area and hypothesize that additional clustering (perhaps due to friendship structure) and duration distribution may be important effects to include. Among subsets of our model that we considered, Model 4 (which also includes architectural, demographic, and organizational data) does the best job of reproducing outbreak size, and its predictions are similar to those based on our full model. Model 4 differs only from the full model in its exclusion of sociality effects for research groups. Since these were not significant in the full model, we, in fact, do not have evidence that these effects are nonzero, and we consider Model 4 to be the best-fitting model. Our analyses of Kendall's τ show that both the full model and Model 4 capture a relevant part of the individual epidemiological importance of individuals, measured as the expected number of secondary cases they would generate if infected. While they are omitting a lot of information, the vastly reduced models 1–3 still capture some information about the individual importance

of individuals. Hence, it might be worth incorporating easily accessible information about the organizational and architectural structure in future analyses and modeling efforts (see discussions in Smieszek & Salathé (2013) and Chowell & Viboud (2013)).

Our work contains several limitations. First, not everyone responded to the survey. In particular, no professors replied, so we have a biased subset of network actors. It is likely that the contact patterns of professors differ from those represented in our sample, but we have no information regarding how they may differ. It is therefore best to consider our estimates to apply to the network of graduate students, postdocs, and administrative staff, rather than representing the entire contact network within the research institute.

Next, although we adjusted for errors in reporting whether contact occurred or not, we did not adjust for inconsistencies in duration reports. Instead, we made the simplifying assumption that if two participants reported different durations for the same contact, the longer duration is the correct report. Our model could be extended to model actual rather than observed durations, for example, by including and estimating a probability that the two reported durations are within a specified threshold. This would be an interesting avenue for future work, as there was a fair amount of inconsistency in duration reports in the data set (detailed in Smieszek et al. (2012)). We do not have information to estimate a tendency to underestimate or overestimate contact durations. One way to collect such data might be studies that jointly collect contact diary data and wireless sensor data detecting when subjects are in close proximity and face-to-face (such as radio frequency identification data). The first and (to our knowledge), only such study was done by Smieszek et al. (2014), who found that survey reporting was reasonably accurate for long contacts, but highly inaccurate for short contacts. However, they found additional unexplained differences between survey and sensor data, and recommend further research into the cause of these. Another recent paper compares online survey reports (rather than contact diaries) with wireless sensor data and direct observations, and finds major differences in network structures and interaction patterns captured by each of the methods (Sailer et al., 2013).

We made the assumption that no contacts were fabricated. We believe that false positive reports are much less likely to occur than false negatives, but integrating a probability of fabrication into our model is one possible direction for future research. To do so, we would need an additional source of data (e.g. sensor data or observations) to judge whether inconsistent contact reports were due to fabrication or forgetting.

An additional assumption we made is the independence in contact between dyads, conditional on the effects in our model. Further research could explore the effect of higher-order dependencies in contact patterns, such as a transitivity effect. In the ERGM framework such an effect could be naively modeled with an inclusion of a triangles statistic, capturing the increase in the odds of contact between two people who contact the same third person, controlling for other effects in the model. It is modeled more realistically by parametric terms which capture decreasing marginal returns on the number of mutual contacts on the increase in odds of contact (Hunter, 2007). Including such an effect would be quite complicated in this setting. We did perform goodness of fit diagnostics comparing our ERGM fit to the network in which a contact was assumed to occur if reported by at least one of the members,

to a model identical to ours but which adds one such transitivity term. These are included in the supplementary material and indicate that our model captured a good part of transitivity present in the network, but could be improved with an additional transitivity term. However, a previous study modeling social networks for influenza transmission in a high school found that inclusion of mixing preferences and heterogeneity in contact duration was sufficient to capture the level of clustering relevant to disease transmission, and transitivity terms did not increase predictive power of the model (Potter et al., 2012). For this reason, we hypothesize that including network dependency terms would not improve our model for the purpose of disease prediction. However, we suspect that we have omitted some group mixing terms or individual-level predictors from our model which may be important to explain the clustering in our network.

In conclusion, we have used organizational structure, architectural structure, and demographic information to estimate the contact network in a research institute. We adjusted for reporting errors and investigated the impact of this adjustment on network effect estimates. We found the adjustment in this context to have a negligible impact on estimates, although we believe the impact may be substantial in networks in other settings. We found angular distance, workgroup membership, and project collaboration to be highly predictive of contact in this setting. Our epidemic simulation study shows that the network structures we have modeled are important for epidemic predictions and produce different forecasts than a random mixing assumption. Our findings indicate that incorporating architectural and organizational data into large-scale epidemic forecasting models may improve the accuracy of epidemic predictions, thus improving our ability to contain and control epidemics.

Supplementary materials

For supplementary material for this article, please visit http://dx.doi.org/10.1017/nws.2015.22

References

Adams, J., & Moody, J. (2007). To tell the truth: Measuring concordance in multiply reported network data. *Social Networks*, **29**(1), 44–58.

Allen, T. J., & Fustfeld, A. R. (1975). Research laboratory architecture and the structuring of communications. *R & D Management*, **5**(2), 153–64.

Almquist, Z. W. (2012). Random errors in egocentric networks. *Social Networks*, **34**(4), 493–505.

Butts, C. T. (2003). Network inference, error, and informant (in)accuracy: A bayesian approach. *Social Networks*, **25**(2), 103–140.

Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J.-F., & Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, **5**(7), e11596.

Chao, D. L., Halloran, M. E., Obenchain, V. J., & Longini Jr., I. M. (2010). FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Computational Biology*, **6**(1), e1000656.

Chowell, G., & Viboud, C. (2013). A practical method to target individuals for outbreak detection and control. *BMC Medicine*, **11**(1), 1–3.

Duerr, H.-P., Schwehm, M., Leary, C. C., De Vlas, S. J., & Eichner, M. (2007). The impact of contact structure on infectious disease control: Influenza and antiviral agents. *Epidemiology and Infection*, **135**(7), 1124–1132.

Eames, K. T. D. (2008). Modelling disease spread through random and regular contacts in clustered populations. *Theoretical Population Biology*, **73**(1), 104–111.

Ferguson, N. M., Cummings, D. A. T., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., & Burke, D. S. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, **437**(7056), 209–214.

Ferguson, N. M., Cummings, D. A. T., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature*, **442**, 448–452.

Geyer, C. J. (2009). *Trust: Trust region optimization*. R package version 0.1-2.

Gilbert, P. (2011). *Numderiv: Accurate Numerical Derivatives*. R package version 2010.11-1.

Goodreau, S. M., Kitts, J. A., & Morris, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent Social Networks. *Demography*, **46**(1), 103–125.

Groendyke, C., Welch, D., & Hunter, D. R. (2012). A network-based analysis of the 1861 hagelloch measles data. *Biometrics*, **68**(3), 755–765.

Halloran, M. E., Ferguson, N. M., Eubank, S., Longini Jr., I. M., Cummings, D. A. T., Lewis, B., . . . Cooley, P. (2008). Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences*, **105**(12), 4639–4644.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., & Morris, M. (2003). *Statnet: Software tools for the statistical modeling of network data*. The Statnet Project. Retrieved from ⟨http://www.statnet.org⟩, Seattle, WA.

Helleringer, S., Kohler, H.-P., Kalilani-Phiri, L., Mkandawire, J., & Armbruster, B. (2011). The reliability of sexual partnership histories: Implications for the measurement of partnership concurrency during surveys. *AIDS*, **25**(4), 503.

Herfst, S., Schrauwen, E. J. A., Linster, M., Chutinimitkul, S., de Wit, E., Munster, V. J., . . . Fouchier, R. A. M. (2012). Airborne transmission of influenza A/H5N1 virus between ferrets. *Science*, **336**(6088), 1534–1541.

Hillier, B. (1996). *Space is the machine. A configurational theory of architecture*. Cambridge University Press.

Hillier, B., & Hanson, J. (1984). *The social logic of space*. Cambridge University Press.

Hillier, B., & Iida, S. (2005). Network effects and psychological effects: A theory of urban movement. In A. van Nes (Ed.), *Proceedings of the 5th International Space Syntax Symposium* (pp. 551–64). Techne Press.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, **100**(469), 286–295.

Hornbeck, T., Naylor, D., Segre, A. M, Thomas, G., Herman, T., & Polgreen, P. M. (2012). Using sensor networks to study the effect of peripatetic healthcare workers on the spread of hospital-associated infections. *Journal of Infectious Diseases*, **206**(10), 1549–1557.

Hunter, D. R. (2007). Curved exponential family models for social networks. *Social Networks*, **29**(2), 216–230.

Hunter, D. R., & Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, **15**(3), 565–583.

Iida, S. (2009). Segmen reference manual. *1.0d, version 0.65*. Bartlett School of Graduate Studies.

Iozzi, F., Trusiano, F., Chinazzi, M., Billari, F. C., Zagheni, E., Merler, S., . . . Manfredi, P. (2010). Little Italy: An agent-based approach to the estimation of contact patterns- fitting predicted matrices to serological data. *PLoS Computational Biology*, **6**(12), e1001021.

Isella, L., Romano, M., Barrat, A., Cattuto, C., Colizza, V., Van den Broeck, W., . . . Tozzi, A. E. (2011). Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS ONE*, **6**(2), e17144.

Kelso, J., Milne, G., & Kelly, H. (2009). Simulation suggests that rapid activation of social distancing can arrest epidemic development due to a novel strain of influenza. *BMC Public Health*, **9**(1), 117.

Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electronic Journal of Statistics*, **6**, 1100–1128.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models (monographs on statistics and applied probability 37)*. London: Chapman Hall.

Milne, G. J., Kelso, J. K., Kelly, H. A., Huband, S. T., & McVernon, J. (2008). A small community model for the transmission of infectious diseases: Comparison of school closure as an intervention in individual-based models of an influenza pandemic. *PLoS ONE*, **3**(12), e4005.

Moser, M. R., Bender, T. R., Margolis, H. S., Noble, G. R., Kendal, A. P., & Ritter, D. G. (1979). An outbreak of influenza aboard a commercial airliner. *American Journal of Epidemiology*, **110**(1), 1–6.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., . . . Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, **5**(3), e74.

Noether, G. E. (1981). Why kendall tau. *Teaching Statistics*, **3**(2), 41–43.

Potter, G. E., & Hens, N. (2013). A penalized likelihood approach to estimate within-household contact networks from egocentric data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**(4), 629–648.

Potter, G. E., Handcock, M. S., Longini, I. M., & Halloran, M. E. (2011). Modeling within-household contact networks from egocentric data. *The Annals of Applied Statistics*, **5**(3), 1816–1838.

Potter, G. E., Handcock, M. S., Longini, I. M., & Halloran, M. E. (2012). Estimating within-school contact networks to understand influenza transmission. *The Annals of Applied Statistics*, **6**(1), 1–26.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Read, J. M., Edmunds, W. J., Riley, S., Lessler, J., & Cummings, D. A. T. (2012). Close encounters of the infectious kind: Methods to measure social mixing behaviour. *Epidemiology and Infection*, **140**(12), 2117.

Russell, C. A., Fonville, J. M., Brown, A. E. X., Burke, D. F., Smith, D. L., James, S. L., . . . & Smith, D. J. (2012). The potential for respiratory droplet – transmissible A/H5N1 influenza virus to evolve in a mammalian host. *Science*, **336**(6088), 1541–1547.

Sailer, K. (2010). *The space-organisation relationship. On the shape of the relationship between spatial configuration and collective organisational behaviours*. Ph.D. thesis, Technical University of Dresden.

Sailer, K., & McCulloh, I. (2012). Social networks and spatial configuration – how office layouts drive social interaction. *Social Networks*, **34**(1), 47–58.

Sailer, K., & Penn, A. (2009). Spatiality and transpatiality in workplace environments. In D. Koch, L. Marcus, & J. Steen (Eds.), *7th International Space Syntax Symposium* (pp. 095:01–95:11). Royal Institute of Technology KTH.

Sailer, K., Pachilova, R., & Brown, C. (2013). Human versus machine testing validity and insights of manual and automated data gathering methods in complex buildings. Paper presented at *9th International Space Syntax Symposium*, 31 Oct.–3 Nov. 2013.

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.

SMART Research Team. (2013). *SOCIAL MIXING AND RESPIRATORY TRANSMISSION in SCHOOLS (SMART Schools) A Report for the Propel and Canon-McMillan School Districts*. Retrieved from `http://www.smart.pitt.edu/archive/files/Report_to_Schools_FINAL.pdf`.

Smieszek, T., Balmer, M., Hattendorf, J., Axhausen, K. W., Zinsstag, J., & Scholz, R. W. (2011). Reconstructing the 2003/2004 H3N2 influenza epidemic in Switzerland with a spatially explicit, individual-based model. *BMC Infectious Diseases*, **11**(1), 115.

Smieszek, T., Barclay, V. C., Seeni, I., Rainey, J. J., Gao, H., Uzicanin, A., & Salathé, M. (2014). How should social mixing be measured: Comparing web-based survey and sensor-based methods. *BMC Infectious Diseases*, **14**(1), 136.

Smieszek, T., Burri, E. U., Scherzinger, R., & Scholz, R. W. (2012). Collecting close-contact social mixing data with contact diaries: Reporting errors and biases. *Epidemiology and Infection*, **140**(4), 744–752.

Smieszek, T., Fiebig, L., & Scholz, R. W. (2009). Models of epidemics: When contact repetition and clustering should be included. *Theoretical Biology and Medical Modelling*, **6**(1), 11.

Smieszek, T., & Salathé, M. (2013). A low-cost method to assess the epidemiological importance of individuals in controlling infectious disease outbreaks. *BMC Medicine*, **11**(1), 35.

Snijders, T. A. B. (2002). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, **3**(2), 1–40.

Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Colizza, V., Isella, L., . . . Vanhems, P. (2011). Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BMC Medicine*, **9**(1), 87.

Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.-F., . . . Vanhems, P. (2011). High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, **6**(08), e23176.

Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for Social Networks. *Journal of the American Statistical Association*, **85**(409), 204–212.

Stroud, P., Del Valle, S., Sydoriak, S., Riese, J., & Mniszewski, S. (2007). Spatial dynamics of pandemic influenza in a massive artificial society. *Journal of Artificial Societies and Social Simulation*, **10**(4), 9.

Szendroi, B., & Csányi, G. (2004). Polynomial epidemics and clustering in contact networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **271**(Suppl 5), S364–S366.

Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Rgis, C., . . . Voirin, N. (2013). Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS ONE*, **8**(9), e73970.

Wang, D. J., Shi, X., McFarland, D. A., & Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks*, **34**(4), 396–409.

WHO. (2009). *Influenza (seasonal), fact sheet no. 211*. Retrieved from `www.who.int/mediacentre/factsheets/fs211/en/index.html`.

Žnidaršič, A., Ferligoj, A., & Doreian, P. (2012). Non-response in Social Networks: The impact of different non-response treatments on the stability of blockmodels. *Social Networks*, **34**(4), 438–450.