Variable population mixing in disease models

Ewan Colman, Andreas Modlmeier, David Hughes, Shweta Bansal

May 4, 2017

Abstract

How individuals move, interact, and mix in a population is critical to the success of an infectious disease. In this paper we introduce a novel method to quantify the level of mixing in a social system and demonstrate how it contributes to the risk of epidemic outbreak. We develop a dynamic model in which the contact rates between individuals varies heterogeneously across the population. The parameter that controls this heterogeneity, $\phi \geq 0$, also controls the propensity of individuals to mix with other members of the population. By fitting this model to dynamic network data, the mixing parameter can be estimated in various social systems. We use data from food sharing interactions between ants, face-to-face contacts between humans, and online social networks. We examine relationship between the level of mixing and the number of secondary cases caused by an infected individual and find that mixing in a population accelerates the spread of disease.

1 Introduction

Infectious diseases thrive when they have regular and easy access to susceptible victims. The variety of known diseases show us that this can be achieved in a number of ways. Airborne diseases such as influenza prosper when groups of people or livestock congregate in relatively small spaces [1, 2]. Pathogens of this type can reach a large number of individuals but have only a small probability of infecting each one [3]. Sexually transmitted diseases, on the other hand, compensate for their relative scarcity of opportunities by having a high probability of transmission during intercourse and a slow immune response once the host is infected [4].

From these examples it would seem clear that one of two approaches should be applied to model infectious disease. We could treat the system as a well-mixed fluid in which the spread of disease depends solely on the size and density of the population [5, 6]. This is generally known as the mass-action model and is the most established approach to disease modeling [7]. Alternatively we could examine the network of social or sexual relationships and predict how an infection might flow from one node to the next [8].

Now consider an ant colony. Through contact ants are able to transmit parasitic fungi, and through trophallaxis (mouth-to-mouth food sharing) a variety of other infections can spread [9, 10]. Each ant has a role in the society, e.g. forager, nurse etc., and this determines which spatial region in the nest they will occupy [11, 12]. Infectious interactions are therefore driven in part by chance encounters, and in part by preference to be where their designated role requires them to be.

Thus, we cannot say that the system is well-mixed and that interactions occur between any random pair of ants, but at the same time we cannot predict where interactions will occur based on the social structure alone. To model disease spread in this context we need an approach that mediates between the network-based disease models and those built on the well-mixed assumption.

One approach is to consider what happens when edges change over time [13]. As time progresses, the number of new connections one accumulates will grow at a rate that depends on the rate at which the connections change. As we would expect, higher rates of mixing lead to higher rates of disease transmission and an increased risk of an epidemic outbreak [14]. While this version of mixing is clearly relevant to disease outcomes, analysis of this model has yet only considered randomly assigned connections, whereas most problems with traditional models result from the fact that social structure is non-random [15].

In studies of human disease this problem is often answered by incorporating a mixing matrix into the disease model, sometimes referred to as a WAIFW (who acquires infection from whom) matrix [6]. This approach involves partitioning the population into a number of categories, for example by age or gender, then assigning different contact for each different type of interaction [16, 17]. This approach can be useful on a large scale but to be applicable in localized settings, such as at a school or workplace, the detail required in order to categorize each individual is much higher.

In this paper we show that it is possible to analyze mixing in small localized populations without large amounts of personal information, instead we rely on a general theory of human and animal social behavior. We first mention some related literature that motivates our approach to the prob-

lem (Section 2.1). We model the interactive behavior of individuals in a population by considering their propensity to socially mix (Section 2.2). One parameter in the model, ϕ, controls the heterogeneity of contact probabilities and we show how it can be tuned to achieve the levels of mixing observed in various human and animal populations (Section 2.3).

We then consider how mixing affects the spread of disease (Section 2.4). We derive a formula for the expected number of secondary infections originating from one infected host as a function of ϕ (for a disease with susceptible $\rightarrow$ infectious $\rightarrow$ removed dynamics), estimate ϕ in 25 dynamic contact networks, and show that mixing is directly related to the number of secondary infections caused (Section 2 and in Figures 1 and 2).

2 Modeling and analysis

2.1 Background

All social species are driven by an intrinsic need to bond with others. In most cases, the number of social contacts an individual can maintain is constrained by the limitations of time and energy [reference?]. In primates, including humans, it is also restricted by the cognitive capacity of the species [18, 19, 20]. This means individuals are frequently confronted with a choice: to invest in forming new relationships or to invest in maintaining pre-existing ones. Since the second option is usually favored over the first, social networks change at a relatively slow pace.

Another result of this is that the distribution of relationship strengths for most people is heterogeneous; most time is invested in close contacts (best friends, family members, etc.), less is invested in the wider friendship circle, and as the circle extends to a wider selection of people, the frequency of interaction decreases [21, 18]. We see the effect of this heterogeneity when we observe an individual over the course of a day (or week, or year) and compare their overall amount of social activity to the number of social connections they made during that time [22]. Gregarious individuals, for example, are those who engage in a large number of interactions and also have a relatively large set of social contacts. Compared to others they show little preference to interact with one individual over any other, or in other words, they have low heterogeneity in the distribution of their relationship strengths. By answering the question of how such social choices are made, the level of sociality can be modeled and quantified [23].

Since the question of social preference is based only decision making, it can be asked without the need to consider the time-scale on which the social activity occurs. We must, however, be conscious of the role of time in how an interaction is interpreted in the context of transmissible disease. The assumption [choose a better word EC] that we make, and continue to make throughout this paper, is that all interactions carry the same probability of infection β. We are thus disregarding the fact that interactions vary in duration, intimacy, and contact type; β represents a probability of infection that takes all such variables into account.

Thus, each interaction between an infected host and a susceptible contact presents a new opportunity for the disease to transmit. Given this interpretation we can deduce that restrained social behavior helps to mitigate the risks of disease spread; once a disease has spread from one individual to another, repetition of the same interaction presents no advantage for the infectious agent (who would rather its host be interacting with someone unfamiliar and susceptible). Repeated contact with a small number of individuals is therefore, in general, a safer strategy than constantly seeking to interact with less familiar members of the population [23].

2.2 Social behavior

Our analysis concerns a closed system containing a set $\mathcal{N}$ of N individuals who can interact with each other in some way. The model has only one tunable parameter which can be interpreted as both the heterogeneity of relationship strengths, and the level of mixing in the population (or population fluidity). By using maximum likelihood methods we are able to measure and compare this quantity across all of our data.

We start by considering one focal individual i and its relationship to another individual j. Suppose that i is involved in a pairwise interaction. We define the relationship strength $x_{j|i}$, for all $j \in \mathcal{N} \setminus \{i\}$, as the probability that the interaction will be with j.

If at least one interaction has occurred between i and j then we say that an edge exist between them. The probability that this is the case after t interactions is

$$P_{i \rightarrow j}(t) = 1 - (1 - x_{j|i})^t. \tag{1}$$

In order to relate our approach to network theory lets first suppose that a latent network structure exists and interactions only occur between nodes connected in this network. Let k_i be the latent degree of i, i.e. each i has a neighborhood $\mathcal{N}_i$ of k_i neighbors that they interact with at equal rates, then $x_{j|i} = 1/k_i$ if $j \in \mathcal{N}_i$ and 0 otherwise. The observed degree, d_i, increases as t increases, its expectation is $\mathbb{E}[d_i(t)] = k_i[1 - (1 - 1/k_i)^t]$. The well-mixed case is recovered when $\mathcal{N}_i = \mathcal{N} \setminus \{i\}$ and therefore $k_i = N - 1$. This example does not capture the whole range of values that $x_{j|i}$ can potentially take, nor does it give a good fit to any of the data that was used for this study.

We now introduce heterogeneity into the distribution of relationship strengths to create a versatile tunable model that can be fitted to various data sources. In this model we make no assumptions about the relationship between i and j other than that $x_{j|i}$ is drawn from some distribution $\rho(x)$. The probability that an edge exists between i and any node in the network is

$$P_i(t) = 1 - \int \rho(x)(1 - x)^t dx. \tag{2}$$

Letting d_i be the degree of i, the expectation is simply $\mathbb{E}(d_i) = (N-1)P_i(t)$. For a given distribution (ρ) of relationship strengths in a population we now have a formula that connects the number of interactions to the degree.
Our goal is to choose the distribution ρ that produces an accurate recreation of the behavior seen in real social systems. We therefore choose the truncated power law,

$$\rho(x) = \frac{\phi\epsilon^{\phi}}{1-\epsilon^{\phi}} x^{-(1+\phi)} \text{ for } \epsilon < x < 1. \tag{3}$$

The reason for this choosing a power law is that it allows the heterogeneity of the relationship strengths to be controlled by a single parameter ϕ making it adaptable to a wide variety of social systems. The distribution is truncated at ϵ so as not to include an asymptote at $x = 0$. It is truncated 1 to ensure that all values of $x_{j|i}$, which are probabilities, are less than 1.
The value of ϵ is determined by the choice of ϕ. To find ϵ, consider that interactions are pairwise; when i interacts, exactly one other individual is involved. Hence, the expectation of the sum of the $x_{j|i}$'s over all $j \in \mathcal{N} \setminus \{i\}$ is equal to 1. Another way to express this is

$$(N-1)\langle x \rangle = 1 \tag{4}$$

where $\langle x \rangle$ denotes the mean of the distribution $\rho(x)$, and is

$$\langle x \rangle = \frac{\phi\epsilon^{\phi}(1-\epsilon^{1-\phi})}{(1-\phi)(1-\epsilon^{\phi})}. \tag{5}$$

Combining Eq.(4) and Eq.(5) we find that the only possible choice of ϵ is the solution of

$$(A+1)\epsilon^{\phi} - \epsilon - A = 0 \tag{6}$$

where $A = (1-\phi)/(N-1)\phi$. The solution of Eq.(2) is

$$P_i(t) = 1 - \frac{\phi\epsilon^{\phi}(1-\epsilon)^{t+1}}{(1-\epsilon^{\phi})(t+1)} {}_2F_1(t+1, 1+\phi, t+2, 1-\epsilon) \tag{7}$$

The notation ${}_2F_1$ refers to the Gauss hypergeometric function [24]. Recall that $P_i(t)$ is the probability that an edge exists from i to j, for any $j \in \mathcal{N} \setminus \{i\}$, after i has been involved in t interactions. The existence of any edge is therefore determined by a Bernoulli trial independent of the existence of any other. After t interactions the degree of i should therefore follow a binomial distribution $d_i(t) \sim B(N-1, P_i(t))$, however, this gives non-zero probabilities for cases where $d > t$. This only occurs for $0 < t < N$ so we replace the formula in this region with a binomial distribution with the same mean, $(N-1)P_i(t)$, but bounded by t. Thus

$$d_i(t) \sim \begin{cases} B\left(t, \dfrac{(N-1)P_i(t)}{t}\right) & \text{if } 0 < t < N \\ B\left(N-1, P_i(t)\right)) & \text{if } t \geq N \end{cases} \tag{8}$$

2.3 Fitting to data

The previous analysis is used to find the value of ϕ that give the best fit to the data. The model we have described is based on the assumption that the system is closed; over some sampling period the N individuals only interact with others from the same population. For each node i we need to know know the number of time they interacted, t_i, and the number of others they interacted with, d_i. We write this as two vectors $\mathbf{d} = \{d_1, d_2, ..., d_N\}$ and $\mathbf{t} = \{t_1, t_2, ..., t_N\}$.
The family of distributions in Eq.(8) allow us to calculate $P(d|t)$, the probability that an individual will have degree d given that they have interacted t times, for any value of the global parameter ϕ. The log-likelihood function is

$$\log \mathcal{L}(\phi|\mathbf{d}, \mathbf{t}) = \sum_{i=1}^{N} \log[P(d_i|t_i)]. \tag{9}$$

We then compute the maximum likely estimate of ϕ, $\phi = \text{argmax}_{\phi} \log \mathcal{L}(\phi|\mathbf{d}, \mathbf{t})$. Standard error SE_{ϕ} is calculated at 95% confidence intervals using $SE_{\phi} = 1.96/\sqrt{-N(\log \mathcal{L})''}$ where the derivatives of $\log \mathcal{L}$ are computed numerically. The standard error is a measure of confidence that our chosen ϕ is in fact the best choice. It does not tell us how well the model fits the data in the first place. To quantify this we introduce a measure of model fidelity.
To measure model fidelity we compare the likelihood of the proposed model it to a null model that represents the most random, i.e. uniformly distributed, possible degree distribution for each given t. The null model equivalent of Eq.(8) is

$$d_i(t) \sim \begin{cases} U(0, t) & \text{if } 0 < t < N \\ U(0, N-1) & \text{if } t \geq N. \end{cases} \tag{10}$$

Model fidelity, f_{ϕ}, quantifies the amount to which the proposed model fits the data when compared to an equivalent null model. We define it as

$$f_{\phi} = (1/N)[\log \mathcal{L}(\phi|\mathbf{d}, \mathbf{t}) - \log \mathcal{L}(\text{null}|\mathbf{d}, \mathbf{t})]. \tag{11}$$

Because we are using observed values of t_i this approach controls for fact activity levels may vary between data sets.

2.4 Disease transmission

Modeling the effect of ϕ on the number of secondary infections We derive the expectation of the number of secondary infections caused by one infectious individual in the population. We shall refer to that individual as i. We will assume that the disease in question follows SIR dynamics according to the following parameters:

Table 1: Summary of parameters and variables

Social behavior		Disease transmission	
N	Number of nodes	β	The probability of transmission given that contact has occurred
$\sum t_i$	The total number of interactions of all nodes	γ	Recovery rate of the disease model. Chosen so that the mean number of infectious contacts is the same across all data-sets Eq.(19)
ϕ	The mixing parameter. The optimal value calculated from the process described in Section 2.2	$\bar{r}$	Mean individual reproduction number based on disease simulation.
ϵ	The lower cut-off for the relationship strength distribution, Eq.(3)	SE_r	Standard error of the reproduction number based on disease simulation
SE_ϕ	The standard error of the estimate of ϕ	$\|e\|$	Absolute error. Sum of the differences between r_i predicted by Eq.(15) and r_i simulated
f_ϕ	Model fidelity. Given by Eq.(11)		

a_i	Contact rate of i
ϕ	Mixing parameter
γ	Recovery rate
β	Transmission probability

If i is infectious for a time period of length τ then the probability that the infection has spread from i to any other individual j is equal to the probability that at least one contact that occurred in that time was infectious. Since contact between individuals follows a Poisson process with rate a_i, and each contact is infectious with probability β, this is

$$T_{i\to j}(\tau, a_i, x_{j|i}) = 1 - \exp(a_i x_{j|i} \beta \tau) \tag{12}$$

As in Section 2.2 we make no assumptions about the relationship between i and j other than that $x_{j|i}$ is drawn from the distribution given by Eq.(3). The probability that transmission occurs from i to any other node in the network is

$$\begin{aligned} T_i(\tau, a_i) &= \int_0^\infty \rho(x) T_{i\to j}(\tau, a_i, x) dx \\ &= 1 - \frac{\phi\epsilon^\phi}{1-\epsilon^\phi} \int_\epsilon^1 x^{-(1+\phi)} \exp(a_i x_{j|i} \beta \tau) dx \\ &= 1 - \phi \sum_{k=0}^{\infty} \frac{(-a_i\beta\tau)^k}{(k-\phi)k!} \frac{\epsilon^\phi - \epsilon^k}{1-\epsilon^\phi}. \end{aligned} \tag{13}$$

Since the rate at which infected individuals move from the infected state (I) to the recovered state (R) is constant, the time spent in the infectious period follows an exponential distribution. In other words the probability that an individual spend a duration τ in the infectious state is $\gamma \exp(-\gamma\tau)$. Integrating Eq.(13) across all possible values of τ we get

$$\begin{aligned} T_i(a_i) &= \int_0^\infty \gamma e^{-\gamma\tau} T_i(\tau, a_i) d\tau \\ &= 1 - \phi \sum_{k=0}^{\infty} \frac{(-a_i\beta/\gamma)^k}{k-\phi} \frac{\epsilon^\phi - \epsilon^k}{1-\epsilon^\phi}. \end{aligned} \tag{14}$$

The quantity T_i is the probability that i will infect j for any $j \in \mathcal{N} \setminus \{i\}$. To get the expected number of secondary infections that come from i we simply have to multiply by the number of susceptibles, which is the entire population minus i. We call this the individual reproduction number r_i, not to be confused with the basic reproduction number R_0 which is usually reserved for use as a population level statistic. The individual reproduction number, $r_i(a_i)$, is a function of the rate of contact of the individual. We have that $r_i(a_i) = (N-1)T_i(a_i)$; it is not possible to express $T_i(a_i)$ in terms of N, however, we can express N in terms of ϵ which we know from Eq.(6) decreases as N increases. Using $N - 1 = 1/\langle x \rangle$ and Eq.(5)

$$\begin{aligned} r_i(a_i) &= \frac{T_i(a_i)}{\langle x \rangle} \\ &= \frac{1-\phi}{\phi(\epsilon^\phi - \epsilon)} \left[1 - \epsilon^\phi - \phi \sum_{k=0}^{k=\infty} \frac{(-a_i\beta/\gamma)^k}{k-\phi} (\epsilon^\phi - \epsilon^k) \right] \end{aligned} \tag{15}$$

which can also be expressed using hypergeometric functions

$$\begin{aligned} r_i(a_i) =& \frac{1-\phi}{\phi(\epsilon^\phi - \epsilon)} \left[1 - \epsilon^\phi + \epsilon^\phi {}_2F_1(-\phi, 1, 1-\phi; -a_i\beta/\gamma) \right. \\ & \left. - {}_2F_1(-\phi, 1, 1-\phi; -\epsilon a_i \beta/\gamma) \right] \end{aligned} \tag{16}$$

Noting that the taking the limit in Eq.(15) as $\epsilon \to 0$ is equivalent to $N \to \infty$ we can also say

$$\lim_{N\to\infty} r_i(a_i) = \frac{1-\phi}{\phi} [-1 + {}_2F_1(-\phi, 1, 1-\phi; -a_i\beta/\gamma) \tag{17}$$

if $\phi < 1$ and

$$\lim_{N\to\infty} r_i(a_i) = a_i\beta/\gamma \tag{18}$$

if $\phi > 1$ (at $\phi = 1$, $\rho(x)$ is not defined). Arriving at this solution requires the use of L'Hopital's Rule.

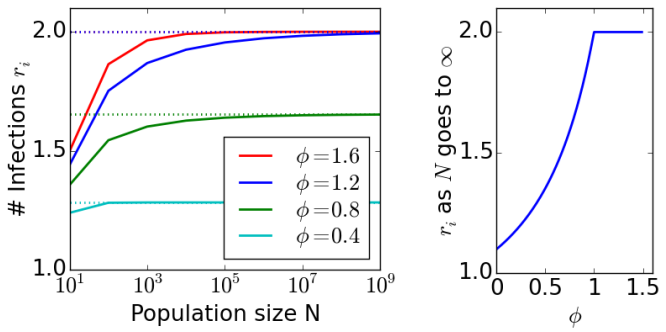

Figure 3: Plot of Eq.(16). As population size increases the expected number of secondary infections converges to the value given by Eqs.(17) and (18). This value increases with ϕ up to $\phi = 1$ and then saturates. The rate of convergence, however, continues to increase.

2.5 Disease simulation

Because the fidelity of the social behavior model, i.e. the extent to which it agrees with the data, varies across the different social settings, we expect that the predictions made in Section 2.4 are only applicable to a some of our data-sets. To test how accurate the prediction of Eq.(16) is, we simulated the effects of transmission on the real contact data.
The collection of data-sets we are comparing is diverse, and social activity happens on dramatically different timescales. To control for this variability the recovery rate γ is adjusted. We choose γ to be

$$\gamma = \frac{2\beta \sum t_i}{N \Delta_t s} \tag{19}$$

where Δ_t is the duration of the time-frame of the data. Eq.(19) is equivalent to choosing γ such that, if the system is well-mixed, then an individual with the mean rate of activity is expected to directly infect s others. In all the results presented we set $s = 2$.
For every individual, i, the simulated reproduction number r_i^{sim} is found by averaging the number of successful infections caused by i over 10^3 simulation trials. Each trial followed the following procedure:

1. A time τ is chosen randomly and uniformly between the beginning and end of the time-frame of the data
2. The length of infectious period Δ_I is generated from an exponential distribution with rate parameter γ
3. A list L of interactions that involved i between time τ and $\tau + \Delta_I$ is generated. If $\tau + \Delta_I$ is beyond the time-frame of the data then interactions from the beginning of the sampling time-frame are used in place of the missing data.
4. Each interaction in the set L is removed with probability $1-\beta$ and r_i is the number of remaining individuals $j \in \mathcal{N} \setminus \{i\}$ that have interactions in L

This gives a reproduction number for every individual in the system. In Table 2 and Figure 2 we provide the mean $\bar{r}$ and standard error SE_r over the population.
Finally, to measure the accuracy of Eq.(16) we calculate the mean absolute error $|e|$. We first calculate the rate of activity $a_i = t_i/\Delta_t$ which, along with the associated values of N, phi, and ϵ, is used in Eq.(16) to compute r_i. The error is given by

$$|e| = \frac{1}{N} \sum_{i \in \mathcal{N}} |r_i - r_i^{\text{sim}}| \tag{20}$$

3 Data

We tested our analysis to 20 study systems. Details of the number of nodes and total number of interactions can be seen in Table 2.

3.1 Ant trophallaxis networks

We collected data from three carpenter ant colonies (Camponotus Pennsylvanicus). In nature, carpenter ant foragers consume liquid food and, upon returning to the nest, regurgitate it into the mouths of their nest-mates, a process known as trophallaxis. Typically, foragers will only give food to a small number of other ants; to feed the entire colony it gets passed through a complex network of feeding interactions [25]. Trophallaxis is also an important form of communication and a way that information about the state of the colony can be shared by all of its members [9, 10].
We placed colonies of approximately 80 ants in a nest designed to replicate the conditions found in nature. The colony was first given a restrictive area of 65×42mm to live (high density) the ants were given several days to adjust before 4 hours of trophallaxis activity was recorded. The nest was then expanded by a factor of 4 (low density) and after another adjustment period another 4 hours were recorded. The process was repeated for 3 unrelated colonies.
[Something about ant diseases]

3.2 Sociopatters human interaction networks

We use human contact data from the Sociopatterns project (sociopatterns.org) [26, 27, 28]. Participants wore radiofrequency identification sensors that detect face-to-face proximity of other participants within 1-1.5 meters in 20-second intervals. Each dataset lists the identities of the people in contact, as well as the 20-second interval of detection. The timing and duration of contacts are known with a resolution of 20-seconds. To exclude contacts detected while participants momentarily walked past one another, only contacts detected in at least two consecutive intervals are considered interactions.
The data we use comes from two studies: an academic conference which occurred over the course of 3 days, which we divided into 3 separate 24-hour data-sets, and a primary school for which there are two days of data, which we

Table 2: A description of each row header is given in table ...

Data set	N	$\sum t_i$	ϕ	$\epsilon \times 10^3$	$SE_\phi \times 10^2$	f_ϕ	$\gamma \times 10^4$	$\bar{r}$	SE_r	$\lvert e \rvert$
ants_1_high	74	496	0.962	2.030	0.931	0.450	1.164	1.478	0.113	0.135
ants_2_high	68	318	0.936	2.105	1.237	0.427	0.812	1.276	0.125	0.231
ants_3_high	76	674	1.107	2.752	0.865	0.655	1.540	1.590	0.143	0.119
ants_1_low	70	606	1.116	3.075	0.962	0.782	1.503	1.589	0.156	0.099
ants_2_low	79	547	0.972	1.926	0.877	0.626	1.202	1.475	0.110	0.124
ants_3_low	83	610	0.957	1.738	0.776	0.536	1.276	1.466	0.138	0.127
conference_0	93	663	0.631	0.356	0.532	0.221	0.206	1.158	0.118	0.188
conference_1	92	650	0.521	0.150	0.561	0.016	0.204	1.137	0.088	0.254
conference_2	84	477	0.288	0.005	0.692	0.424	0.164	0.870	0.070	0.223
hospital_0	51	778	0.614	0.878	0.779	0.769	0.441	1.136	0.141	0.138
hospital_1	49	1075	0.592	0.825	0.741	0.845	0.635	1.136	0.129	0.162
hospital_2	51	855	0.681	1.244	0.774	0.803	0.485	1.252	0.156	0.116
hospital_3	50	743	0.584	0.759	0.802	0.713	0.430	1.136	0.138	0.145
school_0	237	6420	0.620	0.070	0.086	1.245	0.784	1.298	0.041	0.145
school_1	238	6514	0.630	0.076	0.086	1.148	0.792	1.288	0.047	0.150
twitter_1	245	3135	0.763	0.192	0.143	−0.016	0.370	1.317	0.124	0.222
twitter_3	23	292	0.215	0.116	2.553	0.176	0.367	0.736	0.196	0.132
twitter_5	195	3322	0.435	0.010	0.192	0.018	0.493	0.832	0.084	0.234
twitter_6	30	70	0.595	2.021	4.088	−0.322	0.068	0.754	0.241	0.370
twitter_10	24	84	0.528	2.204	4.512	0.184	0.101	0.896	0.209	0.209
twitter_11	72	303	0.321	0.016	0.998	−0.002	0.122	0.821	0.097	0.283
twitter_13	106	1903	0.759	0.607	0.364	0.374	0.519	1.175	0.148	0.228
twitter_15	49	191	0.315	0.047	1.678	0.162	0.113	0.888	0.120	0.250
twitter_16	32	325	0.462	0.796	1.714	0.354	0.294	1.022	0.178	0.131
voles_BHP	195	1339	0.666	0.141	0.222	0.368	132.051	1.385	0.057	0.170
voles_KCS	193	1874	0.719	0.206	0.188	0.632	186.728	1.504	0.059	0.143
voles_PLJ	233	2126	0.743	0.183	0.162	0.448	175.470	1.524	0.055	0.145
voles_ROB	77	381	0.765	0.982	0.876	0.243	95.155	1.288	0.080	0.185
bats_0	16	290	0.427	3.105	3.359	0.640	6.380	0.645	0.072	0.221

divided into 2 24-hour data-sets. [add hospital reference] [26],[28]

3.3 Twitter mentions networks

Mentions on Twitter act as way to send messages directly to particular individuals. Although the message is broadcast to all the followers of the account of the sender it will appear in the inbox of the receiver and is therefore likely to be noticed. Mentions often replied to [22]. We count an interaction as a reciprocated twitter mention, i.e. the motif A→B, B→A is an interaction for A. We ignore all mentions that are not followed by reciprocation i.e. the motif A→B, A→B, B→A, B→A is counted as exactly 1 interaction for A.

In the study from which we take this data [29], communities were first identified before their intensive data collection began. A range of community detection algorithms were used to decide which user accounts belong to the group and which do not, the goal being to find communities such that its members send messages within their community significant more than to user accounts outside the community. The N individuals in the community form their own sub-population which acts somewhat like a closed system, a requisite for applying our analytical methods.

We use the last 24 hours of activity in each of the 17 communities. We discard those for which the number of active users (during the 24 hour period) is either less than 20 or more than 250.

[Something about information transmission in social media]

3.4 Vole contact networks

Data was collected from a population of wild voles (Microtus agrestis) to assess the role of space in determining the structure of social networks [30]. [Sentence about vole behavior and what an interaction is].

In each of four field sites 100 traps were placed in a square grid covering 0.3 hectares. Bait was put into the traps and then three days later observers would check the traps for voles, those who were found were tagged so that they could be recognized should they be caught again. During each trapping session the traps were checked on several consecutive days. If a vole is observed in a trap at any point during

a trapping session then we say that they interacted with any other vole that was observed in the same trap at any point during the same trapping session. The time of the interaction is the day that the trapping session began.
We then discard any voles that had 10 or less interactions and all interactions in which they participated. Since the voles have a very short lifespan and cyclic fluctuations in population size we use only a sub-sample of each data-set. We chose periods of 130 days selected at times of high activity for each of the four experiment sites.
[Something about vole diseases]

3.5 Bat food-sharing networks

Vampire bats share food with each other through regurgitation. In order to initiate such an event a hungry bat will lick the mouth of another bat from whom they hope to receive food. The data we use is a record of mouth-licking observations originally collected to address questions of altruism and reciprocity in bat communities [31, 32].
A population of vampire bats (desmodus rotundus) were kept captive in an enclosure. Out of the 25 bats, 20 were subjected to experimental treatment. In each case, the subject was removed from the enclosure and starved for 24 hours. The observation period of 2 hours began when the starved bat was let back into the enclosure and during this time the usual sources of food were not available. Thus, for the subject bat to feed, interaction with others was necessary. The starvation treatment and observations occurred on a different day for each bat. Some bats were tested more than once so to avoid biasing our results we select only the first day they were tested.
[Something about bat diseases]

4 Results

Humans, ants and online communities exhibit various levels of social mixing. We see from Figure 1 that social connectivity always increases with social activity, however, the rate at which it increases depends on the data. By fitting a stochastic model to each data-set, we demonstrated that this is likely to be a result of the heterogeneity of contact strengths between nodes in the social network, and therefore a direct result of social mixing. The parameter ϕ controls heterogeneity in the model and is used to quantify the amount of mixing between members of the population.
The amount of mixing in a social system relates directly to a higher rate of disease transmission. In an otherwise fully susceptible population, the individual reproduction number r_i, the number of infections caused by an infectious individual i, increases with ϕ. This is known from the results of simulating diseases with SIR dynamics, plotted in Figure 2 (left), and analytically from Eq.(16). For most data-sets the social behavior model does a good job of predicting r_i, however, in cases where the level of agreement between the model and the data is low, so too is its ability to predict disease outcomes (Figure 2 (right)).
As the population size increases to ∞ the individual reproduction number converges for all values of ϕ, however, for populations of infinite size an abrupt phase transition occurs when $\phi = 1$. Below this value ($\phi < 1$) mixing is low and interactions are repeated frequently enough to slow the spread of disease. Beyond this value ($\phi > 1$) the probability that an interaction is repeated is 0 and the number of infections is equal to what it would be in a mass-action model (Figure 3).

5 Discussion

We have measured the level of mixing in a wide variety of social systems and explored how it affects the spread of infectious disease. We have found that disease transmission is influenced by underlying social factors, in particular, the way individuals distribute their attention among other members of the population. By using a power-law to model of heterogeneity in the distribution of contact rates, we are able to accurately reproduce the relationship between the number of interactions of individuals and their degree.
As a tool for predicting the risks of epidemic our analysis has two appealing qualities. Firstly, no information was needed other than two basic quantities associated with social behavior, the degree and the number of interactions of each individual. Our method can therefore be applied without requiring large amounts of personal information. Secondly, The method provides a single numerical value which indicates how mixed a population is, and consequently how easy it will be for an infectious disease to spread.
We have discovered that ants are not particularly averse to the risks of infectious disease. This is not surprising; in addition to the distribution of food, ants use trophallaxis to share information, and share nutrients that are essential for the growth and well-being of the colony. In comparison the costs of infection seem insignificant and the large value of ϕ corroborates this. In the human data, ϕ seems to cluster around 1.6. If this value is indeed indicative of typical human social behavior then we can conclude that the force of infection in the mass-action model is too high. A question which we leave unanswered is how much inaccuracy and over-prediction this leads to when assessments are made about the risks of emerging diseases. We would need consider movement patterns and social behavior over longer durations and greater distances; it should not be assumed that the heterogeneity of contact probabilities we see in localized populations is the correct model at this scale.
Additionally, we should not expect mixing to remain constant across time, particularly as culture changes and social norms shift. If mixing were to increase in terms of sexual contact, for example, a small change in ϕ could be the catalyst for an endemic sexually transmitted disease to become an epidemic. Likewise, as transportation of livestock becomes more efficient, animals can be moved over

larger distances and at greater speed, potentially increasing the mixing and allowing new strains of airborne disease to emerge.

Mixing in an animal population or human society is a variable that can be quantified. Having the ability to summarize large amounts of complexity with one number will improve our ability to understand, and communicate, the risks associated with infectious disease.

References

[1] M. R. MOSER, T. R. BENDER, H. S. MARGOLIS, G. R. NOBLE, A. P. KENDAL, and D. G. RITTER, "An outbreak of influenza aboard a commercial airliner," *American Journal of Epidemiology*, vol. 110, no. 1, p. 1, 1979.

[2] M. Koopmans, B. Wilbrink, M. Conyn, G. Natrop, H. van der Nat, H. Vennema, A. Meijer, J. van Steenbergen, R. Fouchier, A. Osterhaus, *et al.*, "Transmission of h7n7 avian influenza a virus to human beings during a large outbreak in commercial poultry farms in the netherlands," *The Lancet*, vol. 363, no. 9409, pp. 587–593, 2004.

[3] W. E. Bischoff, K. Swett, I. Leng, and T. R. Peters, "Exposure to influenza virus aerosols during routine patient care," *The Journal of Infectious Diseases*, vol. 207, no. 7, p. 1037, 2013.

[4] R. Anderson, S. Gupta, and W. Ng, "The significance of sexual partner contact networks for the transmission dynamics of hiv.," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 3, no. 4, pp. 417–429, 1990.

[5] M. C. de Jong, O. Diekmann, and J. Heesterbeek, "How does transmission of infection depend on population size?," in *Epidemic models: their structure and relation to data*, vol. 5, p. 84, Cambridge University Press, 1995.

[6] R. M. Anderson, R. M. May, and B. Anderson, *Infectious diseases of humans: dynamics and control*, vol. 28. Wiley Online Library, 1992.

[7] M. Begon, M. Bennett, R. G. Bowers, N. P. French, S. Hazel, and J. Turner, "A clarification of transmission terms in host-microparasite models: numbers, densities and areas," *Epidemiology and infection*, vol. 129, no. 01, pp. 147–153, 2002.

[8] R. B. Rothenberg, C. Sterk, K. E. Toomey, J. J. Potterat, D. Johnson, M. Schrader, and S. Hatch, "Using social network and ethnographic tools to evaluate syphilis transmission," *Sexually transmitted diseases*, vol. 25, no. 3, pp. 154–160, 1998.

[9] E. Greenwald, E. Segre, and O. Feinerman, "Ant trophallactic networks: simultaneous measurement of interaction patterns and food dissemination," *Scientific reports*, vol. 5, 2015.

[10] A. C. LeBoeuf, P. Waridel, C. S. Brent, A. N. Gonalves, L. Menin, D. Ortiz, O. Riba-Grognuz, A. Koto, Z. G. Soares, E. Privman, E. A. Miska, R. Benton, and L. Keller, "Oral transfer of chemical cues, growth proteins and hormones in social insects," *eLife*, vol. 5, p. e20375, nov 2016.

[11] J. Gadau, J. Fewell, and E. O. Wilson, *Organization of insect societies: from genome to sociocomplexity*. Harvard University Press, 2009.

[12] D. P. Mersch, A. Crespi, and L. Keller, "Tracking individuals shows spatial fidelity is a key regulator of ant social organization," *Science*, vol. 340, no. 6136, pp. 1090–1093, 2013.

[13] E. Volz and L. A. Meyers, "Susceptible–infected–recovered epidemics in dynamic contact networks," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 274, no. 1628, pp. 2925–2934, 2007.

[14] E. Volz and L. A. Meyers, "Epidemic thresholds in dynamic contact networks," *Journal of the Royal Society Interface*, vol. 6, no. 32, pp. 233–241, 2009.

[15] S. Bansal, B. T. Grenfell, and L. A. Meyers, "When individual behaviour matters: homogeneous and network models in epidemiology," *Journal of The Royal Society Interface*, vol. 4, no. 16, pp. 879–891, 2007.

[16] C. Castillo-Chavez, H. W. Hethcote, V. Andreasen, S. A. Levin, and W. M. Liu, "Epidemiological models with age structure, proportionate mixing, and cross-immunity," *Journal of mathematical biology*, vol. 27, no. 3, pp. 233–258, 1989.

[17] S. D. Valle, J. Hyman, H. Hethcote, and S. Eubank, "Mixing patterns between age groups in social networks," *Social Networks*, vol. 29, no. 4, pp. 539 – 554, 2007.

[18] R. Dunbar, "The social brain hypothesis," *brain*, vol. 9, no. 10, pp. 178–190, 1998.

[19] J. Holt-Lunstad, T. B. Smith, and J. B. Layton, "Social relationships and mortality risk: A meta-analytic review," *PLOS Medicine*, vol. 7, pp. 1–1, 07 2010.

[20] R. Dunbar, *Human evolution: A Pelican introduction*. Penguin UK, 2014.

[21] P. M. Carron, K. Kaski, and R. Dunbar, "Calling dunbar's numbers," *Social Networks*, vol. 47, pp. 151 – 155, 2016.

[22] B. Gonalves, N. Perra, and A. Vespignani, "Modeling users' activity on twitter networks: Validation of dunbar's number," *PLOS ONE*, vol. 6, pp. 1–5, 08 2011.

[23] M. Karsai, N. Perra, and A. Vespignani, "Time varying networks and the weakness of strong ties," *Scientific Reports*, vol. 4, p. 4001, 2014.

[24] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. New York: Dover, 1975.

[25] L. E. Quevillon, E. M. Hanks, S. Bansal, and D. P. Hughes, "Social, spatial, and temporal organization in a complex insect society," *Scientific reports*, vol. 5, 2015.

[26] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, "What's in a crowd? analysis of face-to-face behavioral networks," *Journal of theoretical biology*, vol. 271, no. 1, pp. 166–180, 2011.

[27] P. Vanhems, A. Barrat, C. Cattuto, J.-F. Pinton, N. Khanafer, C. R?gis, B.-a. Kim, B. Comte, and N. Voirin, "Estimating potential infection transmission routes in hospital wards using wearable proximity sensors," *PLoS ONE*, vol. 8, p. e73970, 09 2013.

[28] J. Stehl, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J. Pinton, M. Quaggiotto, W. Van den Broeck, C. Rgis, B. Lina, and P. Vanhems, "High-resolution measurements of face-to-face contact patterns in a primary school," *PLOS ONE*, vol. 6, p. e23176, 08 2011.

[29] N. Charlton, C. Singleton, and D. V. Greetham, "In the mood: the dynamics of collective sentiments on twitter," *Royal Society Open Science*, vol. 3, no. 6, 2016.

[30] S. Davis, B. Abbasi, S. Shah, S. Telfer, and M. Begon, "Spatial analyses of wildlife contact networks," *Journal of The Royal Society Interface*, vol. 12, no. 102, 2014.

[31] G. G. Carter and G. S. Wilkinson, "Food sharing in vampire bats: reciprocal help predicts donations more than relatedness or harassment," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 280, no. 1753, 2013.

[32] G. Carter and G. Wilkinson, "Data from: Food sharing in vampire bats: reciprocal help predicts donations more than relatedness or harassment," 2013.

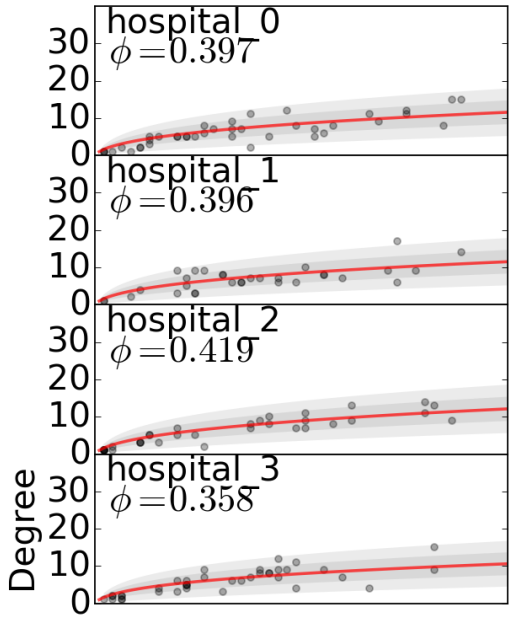

Figure 1: Every data-set used in our analysis as detailed in Section 3. Each point represents one individual in the system. In each case, the mixing parameter ϕ has been tuned to maximize the likelihood of the model using the process described in Section 2.3. The optimal ϕ is given and the curve shows the mean degree of an individual as a function of the number of interactions. The shaded area and the lighter shaded area represent intervals that are one and two standard deviations from the mean respectively. Data points for which the number of interactions is more than 90 are excluded from the figure but not from the inference of ϕ. Reference the formulas that are in the text.

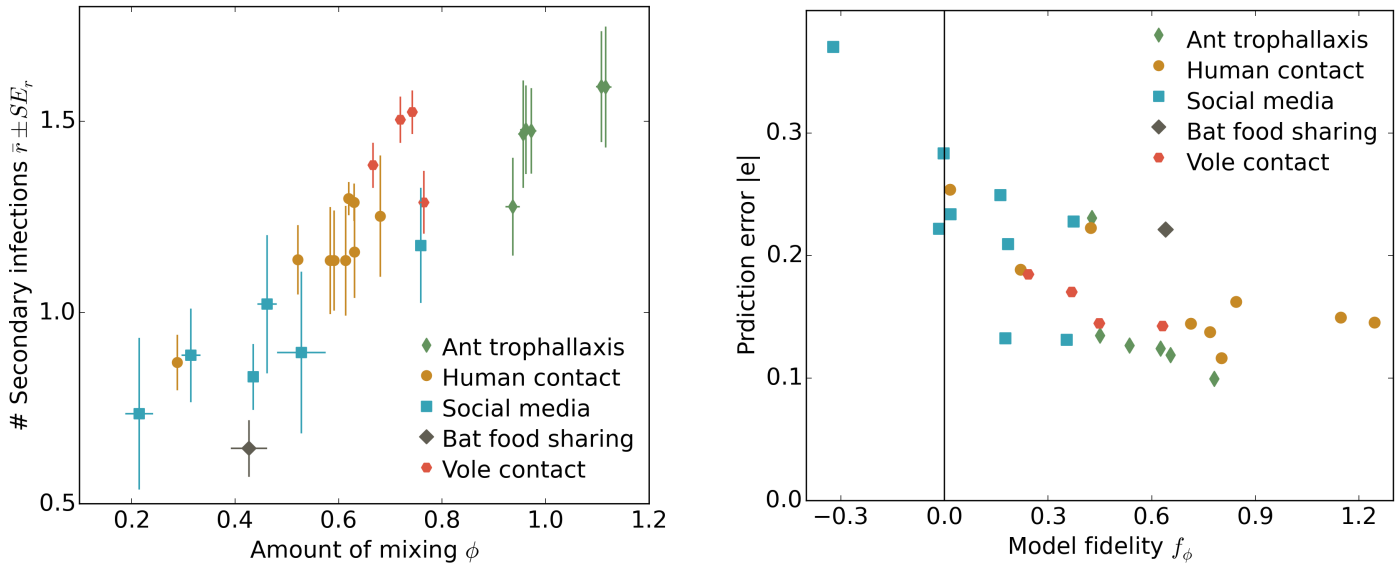

Figure 2: Disease outcomes regarding r_i, the number of secondary infections caused by an individual i. On the left hand side each data point represents one of the systems in the study. The value of the mixing parameter ϕ (as per MLE) with error bars at one standard error, is plotted against the mean r_i over all individuals i in the population (expected values of r_i based on 10^3 simulations) with error bars showing one standard error. On the right hand side each point represents one system in the study. Those which have negative model fidelity are omitted from the plot on the left. As agreement between the model and the data (given by Eq.(11)) increases, so does the agreement between analytical and simulated disease outcomes (given by the total absolute error).