Social fluidity mobilizes contagion in human and animal populations: supporting information

Ewan Colman[*1], **Vittoria Colizza**[2], **Ephraim M. Hanks**[3], **Andreas P. Modlmeier**[4], **David P. Hughes**[4], **and Shweta Bansal**[1]

[1] Department of Biology, Georgetown University, [2] INSERM, Paris, [3] Department of Statistics, Pennsylvania State University, [4] Department of Entomology, Pennsylvania State University

1 Measuring social fluidity

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 we call "social fluidity" and can be interpreted as both the heterogeneity of relationship strengths, and the level of mixing in the population. From the description of the model we derive, analytically, the expected value of the the observed number of interaction partners of an individual, k_i, as a function of the number of times they been observed in an interaction, s_i. By using maximum likelihood methods we are able to measure and compare social fluidity quantity across a wide range of data.

1.1 Mathematical model of social behavior. We start by considering one focal individual i and its relationship to another individual j. Suppose that i is observed interacting with one other individual. We use $x_{j|i}$, for all $j \in \mathcal{N} \setminus \{i\}$, to denote the probability that the interaction will be with j. If at least one interaction has been observed between i and j then we say that an edge exists between them. The probability that this is the case after i has been observed s_i times is

$$\Psi_{i\to j}(s_i) = 1 - (1 - x_{j|i})^{s_i}. \tag{1}$$

When we consider the entire system, $x_{j|i}$ must satisfy two conditions: first, since $x_{j|i}$ is a probability conditioned on the fact that i has interacted, we must have $\sum_j x_{j|i}=1$. Second, the probability that i interacts with j at any given time, which can be expressed $a_i x_{j|i}$ where a_i is the rate of activity of i, is equal to the probability that j interacts with i at any time, $a_j x_{i|j}$. These constraints can be expressed as $AX = X^T A$ and $X\mathbf{1} = \mathbf{0}$, where X is a matrix whose i, j entry is -1 if $i = j$ and $x_{j|i}$ otherwise, A is any diagonal matrix whose ith entry is a_i, and $\mathbf{0}$ and $\mathbf{1}$ are column vectors of length N containing only 0 and 1, respectively.

We now introduce heterogeneity into the distribution of relationship strengths. 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 distribution of marginal $x_{j|i}$ values of the joint distribution $P(X)$). The probability that an edge exists between i and any node in the network after s interactions is

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

Letting k_i be the degree of i, the expectation is simply $\mathbb{E}(k_i) = (N - 1)\psi(s_i)$. 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,

$$\rho(x) = \begin{cases} \dfrac{\phi\epsilon^\phi}{1-\epsilon^\phi} x^{-(1+\phi)} & \text{if} \quad \epsilon < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The reason for 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.

1.2 Choosing ϵ. 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

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

In the limit as $N \to \infty$ this constraint in expectation becomes a hard constraint, with $\sum_j x_{j|i} = 1$. In the finite N case the constraint in expectation due to its mathematical tractability. Combining (4) and (5) we find that ϵ is a solution of

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

where $C = (1 - \phi)/(N - 1)\phi$. Since the function on the left hand side of (6) has only one turning point, (6) can have either 0 or 2 solutions. One solution is $\epsilon = 1$ so the other solution is the only possible choice of ϵ. Thus, to compute ϵ for any given values of ϕ and N, we solve (6) numerically using $\epsilon = 0$ as the starting estimate.

1.3 Degree distribution. Substituting (3) into (2) we have, for an individual that has had s interactions, that

$$\Psi(s) = 1 - \frac{\phi\epsilon^\phi}{1 - \epsilon^\phi}\int_\epsilon^1 x^{-(1+\phi)}(1 - x)^s dx. \tag{7}$$

This simplifies to

$$\Psi(s) = 1 - \frac{\phi\epsilon^\phi(1 - \epsilon)^{s+1}}{(1 - \epsilon^\phi)(s + 1)}{}_2F_1(s + 1, 1 + \phi, s + 2, 1 - \epsilon) \tag{8}$$

The notation ${}_2F_1$ refers to the Gauss hypergeometric function [17]. Recall that $\Psi(s_i)$ 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 s_i interactions. The existence of any edge is therefore determined by a Bernoulli trial independent of the existence of any other. After s interactions the degree of any node should therefore follow a binomial distribution $k(s) \sim B(N - 1, \Psi(s))$, however, this gives non-zero probabilities for cases where $k > s$. This only occurs for $0 < s < N$ so we replace the formula in this region with a binomial distribution with the same mean, $(N - 1)\Psi(s)$, but bounded by s. Thus

$$k_i(s) \sim \begin{cases} B\left(s, \dfrac{(N-1)\Psi(s)}{s}\right) & \text{if } 0 < s < N \\ B\left(N - 1, \Psi(s)\right)) & \text{if } s \geq N \end{cases} \tag{9}$$

[*] ec975@georgetown.edu

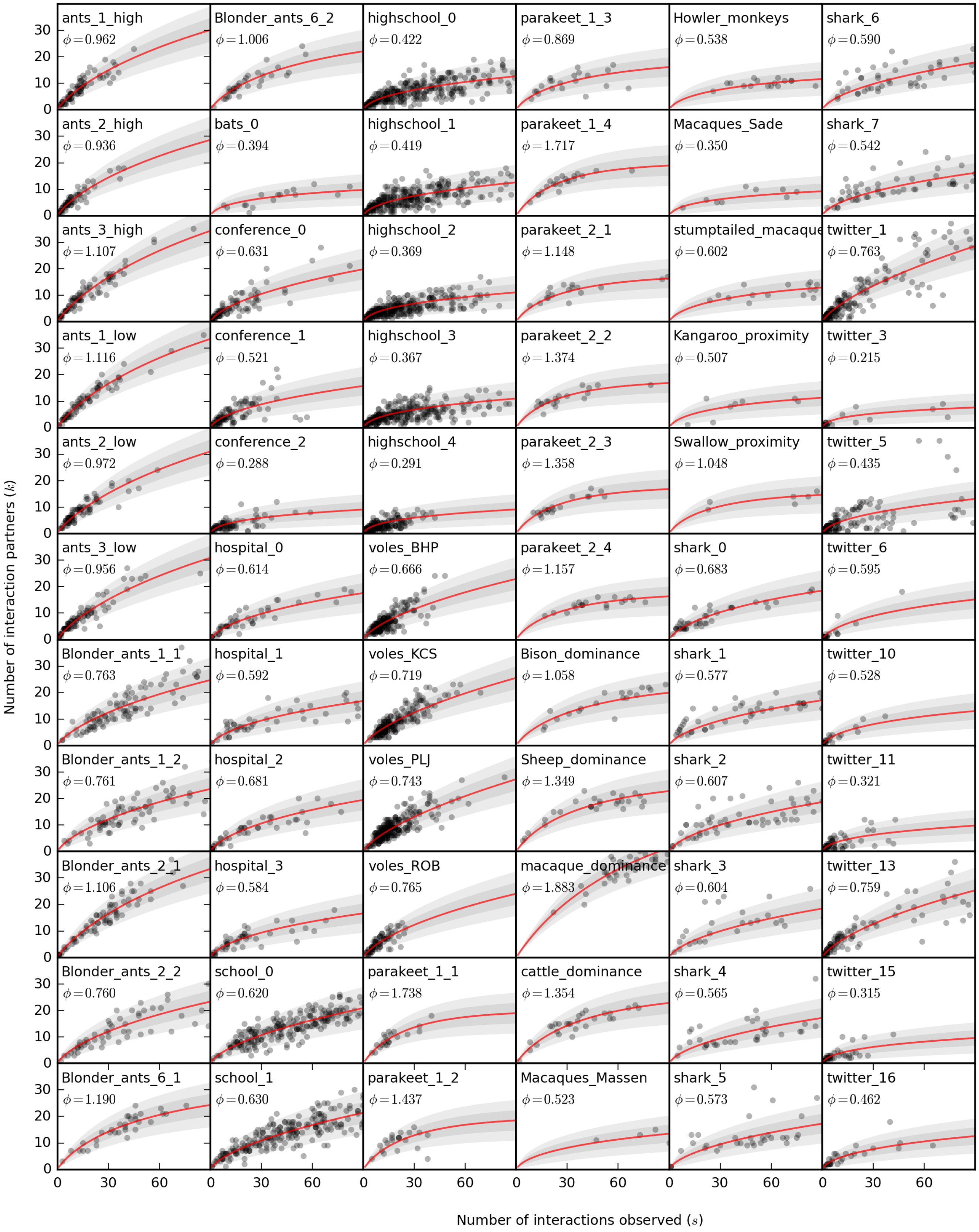

Figure 1: Every data-set used in our analysis as detailed. Each point represents one individual in the system. Motivated by these data, we hypothesize that the form for ρ is power-law. In each case, the parameter ϕ has been tuned to maximize the likelihood of the model. The optimal ϕ is given and the curve shows the expected degree (according to the model) 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 (according to the model) 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.

Since the first part of (9) is an ad-hoc estimate to the true distribution, it is important to clarify that it is consistent with the conditions we would expect for a probability distribution. In particular, we want to show that $(N-1)\Psi(s)/s \leq 1$ in all cases. Consider first that since the least heterogeneous possible situation is when $x_{j|i} = 1/(N-1)$ for all i and j. Putting this into (1) we have

$$\max(\Psi(s)) = 1 - \left(1 - \frac{1}{N-1}\right)^s \tag{10}$$

This gives

$$(N-1)\Psi(s)/s \leq \frac{N-1}{s}\left[1 - \left(1 - \frac{1}{N-1}\right)^s\right]. \tag{11}$$

The value of right hand side hand side decreases with s (since its derivative is negative), therefore the largest possible value of the right hand side occurs when $s = 1$, giving

$$(N-1)\Psi(s)/s \leq 1. \tag{12}$$

1.4 Estimating ϕ in empirical 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 the number of times they interacted, s_i, and the number of partners with whom they interacted, k_i. We write this as two vectors $\mathbf{k} = \{k_1, k_2, ..., k_N\}$ and $\mathbf{s} = \{s_1, s_2, ..., s_N\}$.

The family of distributions in (9) allow us to calculate $P(k|s)$, the probability that an individual will have degree k given that they have interacted s times, for any value of the global parameter ϕ. The log-likelihood function is

$$\log \mathcal{L}(\phi|\mathbf{k}, \mathbf{t}) = \sum_{i=1}^{N} \log[P(k_i|s_i)]. \tag{13}$$

We then compute the maximum likely estimate of ϕ, $\phi = \text{argmax}_\phi \log \mathcal{L}(\phi|\mathbf{k}, \mathbf{s})$. Our estimation is based on the marginal distribution of the degree of each node and does not take into account the interdependencies of the network structure. As such, we do not report standard errors or confidence intervals for our point estimates. For discussion of the use of marginal distributions in likelihood estimation see [60].

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 (9) is

$$k(s) \sim \begin{cases} U(0, s) & \text{if } 0 < s < N \\ \\ U(0, N-1) & \text{if } s \geq N. \end{cases} \tag{14}$$

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

$$f_\phi = (1/N)[\log \mathcal{L}(\phi|\mathbf{k}, \mathbf{s}) - \log \mathcal{L}(\text{null}|\mathbf{k}, \mathbf{s})]. \tag{15}$$

Because we are using observed values of s_i this approach controls for the variation in activity levels between data sets.

Finally, to find a benchmark to compare the model fidelity we measured f_ϕ for several synthetically generated data sets. These data were generated by first selecting 100 values of s between 1 and 100 uniformly at random, then, for each one we select a value of k either from the model distribution given by (9) or from the noise distribution given by (14) ($N = 100$ and $phi = 0.6$). When less than 8% of the synthetic values of k come from the noise distribution, f_ϕ is positive. We conclude that positive values of f_ϕ found in any data-set implies that the model is a good fit to the data.

2 Modelling the spread of disease

The disease model is described as follows: We consider a fully susceptible population of N individuals. At a randomly selected point in time, one individual, i, becomes infectious. They remain infectious for a duration of length τ, where τ is a random variable drawn from an exponential distribution $P(\tau) = \gamma\exp(-\gamma\tau)$. This is equivalent to i being able to recover at any point in time and γ being the probability that recovery occurs at any point during an interval of length 1. The times for which infected individual, i, engages in interaction follows a Poisson process with rate parameter a_i. Given that i is engaged in an interaction, the probability that the interaction is with individual j is $x_{j|i}$. Once contact is established, the probability that the disease will transmit is β.

2.1 The number of secondary infections. We are interested in $r(a_i)$, the expected number of other individuals infected by i, given that the rate of interaction of i is a_i.

The transmission probability, the probability that i infects j during an infectious period of length τ, is equal to the probability that at i makes infectious contact with j at least once during that time. Since the rate of infectious contact between i and j follows a Poisson process with rate $a_i x_{j|i}\beta$, the transmission probability for an infectious period of length τ is derived from the Poisson distribution and is

$$T_{i\to j}(\tau, a_i, x_{j|i}) = 1 - \exp(a_i x_{j|i}\beta\tau). \tag{16}$$

As in Section 1.1 we make no assumptions about the relationship between i and j other than that $x_{j|i}$ is drawn from the distribution given by (3). The probability that transmission occurs from i to any other node in the network is

$$\begin{aligned} T(\tau, a_i) &= \int_0^\infty \rho(x) T_{i\to j}(\tau, a_i, x)dx \\ &= 1 - \frac{\phi\epsilon^\phi}{1-\epsilon^\phi}\int_\epsilon^1 x^{-(1+\phi)}\exp(a_i x_{j|i}\beta\tau)dx \\ &= 1 - \phi\sum_{k=0}^\infty \frac{(-a_i\beta\tau)^k}{(k-\phi)k!}\frac{\epsilon^\phi - \epsilon^k}{1-\epsilon^\phi}. \end{aligned} \tag{17}$$

The probability that i was infectious for a period of duration τ is $\gamma\exp(-\gamma\tau)$. Integrating (17) across all possible values of τ we get

$$\begin{aligned} T(a_i) &= \int_0^\infty \gamma e^{-\gamma\tau} T_i(\tau, a_i)d\tau \\ &= 1 - \phi\sum_{k=0}^\infty \frac{(-a_i\beta/\gamma)^k}{k-\phi}\frac{\epsilon^\phi - \epsilon^k}{1-\epsilon^\phi}. \end{aligned} \tag{18}$$

The quantity T 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. We therefore have that

$$r(a_i) = (N-1)T(a_i). \tag{19}$$

It is not possible to express $T_i(a_i)$ in terms of N so we instead express it in terms of ϵ. By substituting $N-1 = 1/\langle x\rangle$ and (5) into (19) we get

$$\begin{aligned} r(a_i) &= \frac{T_i(a_i)}{\langle x\rangle} \\ &= \frac{1-\phi}{\phi(\epsilon^\phi - \epsilon)}\left[1 - \epsilon^\phi - \phi\sum_{k=0}^\infty \frac{(-a_i\beta/\gamma)^k}{k-\phi}(\epsilon^\phi - \epsilon^k)\right] \end{aligned} \tag{20}$$

which can also be expressed using hypergeometric functions

$$\begin{aligned} r(a_i) =& \frac{1-\phi}{\phi(\epsilon^\phi - \epsilon)}\left[1 - \epsilon^\phi + \epsilon^\phi {}_2F_1(-\phi, 1, 1-\phi; -a_i\beta/\gamma)\right. \\ & \left. - {}_2F_1(-\phi, 1, 1-\phi; -\epsilon a_i\beta/\gamma)\right] \end{aligned} \tag{21}$$

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
$\sum s_i/2$	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 (29)
ϕ	The mixing parameter. The optimal value calculated from the process described in Section 1.1	R_0	Basic reproductive number based on either (24) or disease simulation.
ϵ	The lower cut-off for the relationship strength distribution, (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 (24) and r_i simulated
f_ϕ	Model fidelity. Given by (15)		

2.2 Calibration of time-scales. The collection of data-sets we compare is diverse and social activity happens on different time-scales. Additionally, the types of disease that affect one species are unlikely to affect another. Instead of choosing parameter values that relate to some specific disease, it is more informative to select parameter values for each system separately in a way that exposes the effects of population size and social fluidity. To achieve this the two temporal variables, γ and the mean activity rate, $\langle a_i \rangle$, are calibrated to each other in such a way that R_0 would always be the same value if, hypothetically, the effects of social fluidity and population size were not present.

We define R^* to be the value of R_0 in a large population with homogenous mixing. In this case, the infectious individual, i, will never repeat an interaction with an individual whom they previously infected. In other words, every interaction during the infectious period will, with probability β, cause a new infection, and so R^* can be found by multiplying the mean infectious period by β and the mean rate interactions,

$$R^* = \frac{\beta \langle a_i \rangle}{\gamma}. \tag{22}$$

By estimating the rate parameter a_i from the number of observed interactions s_i as $a_i = s_i/\Delta_t$ where Δ_t is the duration of the time-frame of the data, (22) becomes

$$\gamma = \frac{\beta \langle s \rangle}{\Delta_t R^*}, \tag{23}$$

where $\langle s \rangle$ is the mean of s_i over all individuals. Additionally, by substituting this value of gamma into (21) we get

$$\begin{aligned} r(s) = & \frac{1-\phi}{\phi(\epsilon^\phi - \epsilon)} \left[1 - \epsilon^\phi + \epsilon^\phi {}_2F_1(-\phi, 1, 1-\phi; -R^* s/\langle s \rangle) \right. \\ & \left. - {}_2F_1(-\phi, 1, 1-\phi; -\epsilon R^* s/\langle s \rangle)\right]. \end{aligned} \tag{24}$$

Note that no temporal information appears in this equation. In all the analysis presented we have arbitrarily chosen $R^* = 2$.

2.3 Large population size limit. From section 1.2 we can see that as N grows large, A goes to 0 and the solutions of (6) are $\epsilon = 0$ and $\epsilon = 1$. Expressing R_0 in terms of N is not possible so we instead find the limit as $\epsilon \to 0$. We cannot directly infer $\lim_{\epsilon \to 0} r(a_i)$ from (20) since both the numerator and denominator go to 0. Applying l'Hopital's rule we get

$$r'(a_i) = \frac{(1-\phi)}{(1-\phi\epsilon^{\phi-1})} \left[\epsilon^{\phi-1} + \sum_{k=0}^{\infty} \frac{(-a_i\beta/\gamma)^k}{k-\phi} (\phi\epsilon^{\phi-1} - k\epsilon^{k-1})\right] \tag{25}$$

For $\phi > 1$ we we can find $\lim_{\epsilon \to 0} r'(a_i)$ directly by substituting $\epsilon = 0$,

$$\lim_{N \to \infty} r(a_i) = a_i\beta/\gamma. \tag{26}$$

For $\phi < 1$, we first multiply numerator and denominator by $\epsilon^{1-\phi}$ before substituting $\epsilon = 0$. We get

$$\lim_{N \to \infty} r(a_i)(a_i) = \frac{1-\phi}{\phi} \left[-1 - \sum_{k=0}^{\infty} \frac{(-a_i\beta/\gamma)^k}{k-\phi}\right] \tag{27}$$

which can also be expressed using a hypergeometric function

$$\lim_{N \to \infty} r(a_i)(a_i) = \frac{1-\phi}{\phi} [-1 + {}_2F_1(-\phi, 1, 1-\phi; -a_i\beta/\gamma). \tag{28}$$

(At $\phi = 1$, $\rho(x)$ is not defined).

2.4 Simulating the spread of disease. 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 2.1 are only applicable to a some of our data-sets. To test how accurate the prediction of (24) 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 time-scales. To control for this variability the recovery rate γ is adjusted. We choose γ to be

$$\gamma = \frac{2\beta \sum t_i}{N \Delta_t s} \tag{29}$$

where Δ_t is the duration of the time-frame of the data. (29) 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 t is chosen randomly and uniformly between the beginning and end of the time-frame of the data
2. The length of infectious period τ is generated from an exponential distribution with rate parameter γ
3. A list L of interactions that involved i between time t and $t + \tau$ is generated. If $t + \tau$ 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 3 we provide the mean R_0 and standard error SE_r over the population.

Finally, to measure the accuracy of (21) we calculate the mean absolute error $|e|$. We first calculate the rate of activity $a_i = t_i/\Delta_t$

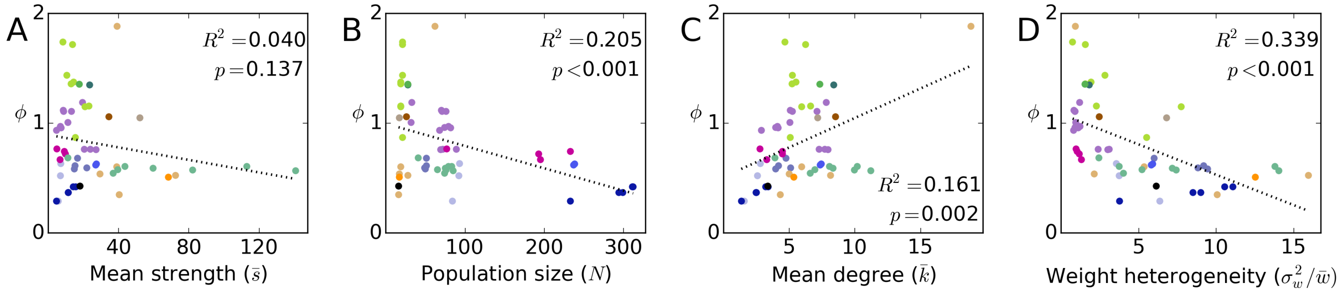

Figure 2: Comparison of social fluidity with other network-based statistics.

which, along with the associated values of N, phi, and ϵ, is used in (21) to compute r_i. The error is given by

$$|e| = \frac{1}{N}\sum_{i\in\mathcal{N}} |r_i - r_i^{\text{sim}}| \qquad (30)$$

3 Sources of empirical data

3.1 Human face-to-face interaction. We use human contact data from the Sociopatterns project (sociopatterns.org). 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 [21], a primary school for which there are 2 days of data [22], a high school which spanned 5 days [23], and a hospital which spanned 4 full days [24]. In each case the data were divided into 24 hour subsets beginning at midnight.

3.2 Ant trophallaxis. We collected data from three carpenter ant *Camponotus pennsylvanicus* colonies. 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 **?**. 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 **??**.

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.

3.3 Ant antennal Contact. The antennae of ants contain highly sensitive olfactory cells. By touching the cuticle of another ant they are able to perceive information (primarily the status of the other ant) which is expressed through hydrocarbons secreted on their cuticle. We use data from [25] which was collected by constant human observation of video footage of ant colonies over a period of approximately 30 minutes. The experiment was performed on 3 unrelated *Temnothorax rugatulus* colonies, each of which was recorded in two sessions separated by a two week period.

3.4 Bat food-sharing. 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 [26].

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.

3.5 Vole territory sharing. Data were collected from a population of wild voles *Microtus agrestis* to assess the role of space in determining the structure of social networks [27]. [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.

3.6 Association by group membership. Since foraging groups and social groups can change from one day to the next, a commonly way to measure the strength of a pairwise relationship is to consider the number of times that the pair were observed in the same group. In most studies of this kind the data is processed in a way that attempts to correct for observation bias and group size heterogeneity, which makes it incompatible for our method of analysis. We were, however, able to use 3 experiments from which the raw data is available. These are kangaroos *Macropus giganteus* [28], barn swallow *Hirundo rustica erythrogaster* [29], howler monkeys *Alouatta palliata* [30], and sharks [31]. The data from each of these papers were collected through intermittent, rather than continuous, observation. We define an interaction as belonging to the same group during one round of observation.

3.7 Grooming. Grooming in monkeys, and other primates, is used to build and social bonds, avoid conflict, and maintain social structures including the dominance hierarchy. We used grooming data from two studies of macaques *Macaca mulatta* [32,33] and one of stumptailed macaques *Macaca arctoides* [34]. Data were collected intermittently rather than through continuous observation. If one animal was grooming another during one round of observations then this would be recorded as a directed interaction. For our analysis we neglect the direction of the interaction (it is unclear whether the direction would have consequences relevant to the spread of disease).

3.8 Aggression and dominance. Aggression between animals can be in the form of a physical fight or a display of dominance that causes one individual to concede. From the literature we obtained aggression data from macaques *Macaca fuscata fuscata* [35], female bighorn sheep *Ovis canadensis* [36], bison *Bison bison* [37], cattle [38], and parakeets *Myiopsitta monachus* [39]. The data from each of these papers were collected during intermittent observation periods. When an animal was determined to be the winner of a dominance encounter then this would be recorded as a directed interaction between the winner and the loser. For our analysis we neglect the direction of the interaction (it is unclear whether the direction would have consequences relevant to the spread of disease).

4 Rejected data

There were numerous studies of animal social behavior that we were not able to adapt for this study because the format of the data was not appropriate. For example, a number of studies publish the results of calculations, such as the half-weight index, rather than the the raw information that they must surely have originally collected. We urge scientists in this field to consider making their raw data available to allow studies such as this one to benefit from their hard work. We mention the following two additional sources of data that were unusable for other reasons.

4.1 Mouse territory sharing. Mice *Mastomys natalensis* [59] were kept in a large enclosed space that contained nine evenly spaced feeding stations. Four feeding stations had sensors that recorded the moment when a mouse passed through its door. Interactions were recorded when two mice pass through the same feeding station within 30 seconds of each other. Since only the interactions that occur at the feeding stations were observed, the networks are sparse, disconnected, and only a small proportion of the entire population were observed interacting. We tested to see whether increasing the time threshold between visits to the feeding station would create higher density networks. Only at threshold values on the order of hours, which we consider biologically unrealistic, did the networks become dense enough to be considered for our analysis.

4.2 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. 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 [40], 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.

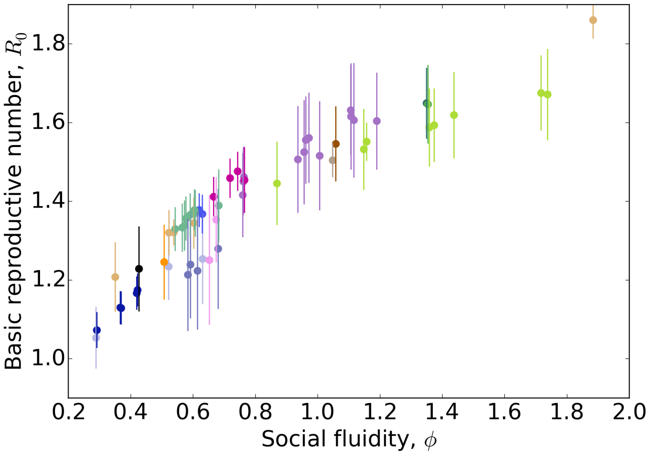

Figure 3: Each point represents a human or animal data-set. For each individual, i, in the population, $r(s_i)$ is calculated from (24) and the mean, R_0 ($\pm$ one standard error), is plotted against its corresponding value of ϕ.

The main reason for not including these data is that they do not relate to biological contagion. However, this electronic communication data did provide a useful comparison to the direct communication data that we did choose to use in our study. We used 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, leaving 9 networks. Model fidelity scores for these 7 networks was fund to be -0.016 (`twitter_1`), 0.176 (`twitter_3`), 0.018 (`twitter_5`), -0.322 (`twitter_6`), 0.184 (`twitter_10`), -0.002 (`twitter_11`), 0.374 (`twitter_13`), 0.162 (`twitter_15`), 0.354 (`twitter_16`).

Table 2: Basic statistics for all networks

Data set	System	Interaction	N	$\sum s_i/2$	$\sum d_i/2$	$\bar{w}$	$\sigma_w^2/\bar{w}$
conference_0	Conference	Face-to-face	93	663	262	2.531	3.571
conference_1	Conference	Face-to-face	92	650	239	2.720	3.742
conference_2	Conference	Face-to-face	84	477	134	3.560	6.367
hospital_0	Hospital	Face-to-face	51	778	213	3.653	9.183
hospital_1	Hospital	Face-to-face	49	1075	250	4.300	8.815
hospital_2	Hospital	Face-to-face	51	855	242	3.533	5.986
hospital_3	Hospital	Face-to-face	50	743	198	3.753	9.439
school_0	Primary school	Face-to-face	237	6420	1744	3.681	5.815
school_1	Primary school	Face-to-face	238	6514	1769	3.682	5.915
highschool_0	High school	Face-to-face	312	4919	1058	4.649	11.072
highschool_1	High school	Face-to-face	311	4419	997	4.432	10.553
highschool_2	High school	Face-to-face	299	3415	746	4.578	8.510
highschool_3	High school	Face-to-face	294	3331	732	4.551	9.010
highschool_4	High school	Face-to-face	233	1084	316	3.430	3.763
ants_1_high	Ant	Food sharing	74	496	293	1.693	0.892
ants_2_high	Ant	Food sharing	68	318	196	1.622	0.799
ants_3_high	Ant	Food sharing	76	674	399	1.689	0.898
ants_1_low	Ant	Food sharing	70	606	356	1.702	0.826
ants_2_low	Ant	Food sharing	79	547	324	1.688	1.220
ants_3_low	Ant	Food sharing	82	610	342	1.784	1.095
Blonder_ants_1_1	Ant	Antennal contact	89	1834	649	2.826	2.399
Blonder_ants_1_2	Ant	Antennal contact	72	1721	556	3.095	3.094
Blonder_ants_2_1	Ant	Antennal contact	71	943	505	1.867	0.904
Blonder_ants_2_2	Ant	Antennal contact	69	1880	549	3.424	3.021
Blonder_ants_6_1	Ant	Antennal contact	33	647	258	2.508	1.185
Blonder_ants_6_2	Ant	Antennal contact	32	362	162	2.235	0.953
bats_0	Bat	Food sharing	16	290	54	5.345	6.131
voles_BHP	Vole	Space sharing	195	1339	640	2.092	1.311
voles_KCS	Vole	Space sharing	193	1874	865	2.166	1.171
voles_PLJ	Vole	Space sharing	233	2126	1048	2.029	1.041
voles_ROB	Vole	Space sharing	77	381	214	1.780	0.962
parakeet_1_1	Parakeet	Aggression	21	175	98	1.733	0.742
parakeet_1_2	Parakeet	Aggression	21	222	110	1.930	2.826
parakeet_1_3	Parakeet	Aggression	21	325	107	2.664	5.518
parakeet_1_4	Parakeet	Aggression	21	291	131	2.094	1.573
parakeet_2_1	Parakeet	Aggression	19	399	113	2.978	7.737
parakeet_2_2	Parakeet	Aggression	19	273	105	2.505	1.931
parakeet_2_3	Parakeet	Aggression	19	247	99	2.330	1.552
parakeet_2_4	Parakeet	Aggression	19	441	126	3.291	2.267
Bison_dominance	Bison	Aggression	26	897	222	2.857	2.467
Sheep_dominance	Sheep	Aggression	28	658	235	2.632	1.806
cattle_dominance	Cattle	Aggression	28	498	205	2.295	1.544
macaque_dominance	Monkey	Aggression	62	2435	1167	2.051	0.908
Macaques_Massen	Monkey	Grooming	28	2025	228	5.853	15.949
Macaques_Sade	Monkey	Grooming	16	647	69	5.829	10.074
stumptailed_macaque	Monkey	Grooming	19	742	113	4.609	5.351
Howler_monkeys	Monkey	Association	17	502	85	2.953	2.135
Kangaroo_proximity	Kangaroo	Association	17	1161	91	6.184	12.532
Swallow_proximity	Swallow	Association	17	886	122	7.262	6.808
shark_0	Shark	Association	41	463	161	2.876	2.427
shark_1	Shark	Association	65	2571	452	5.688	5.207
shark_2	Shark	Association	83	3619	670	5.401	3.749
shark_3	Shark	Association	79	8926	805	11.088	13.778
shark_4	Shark	Association	85	11955	953	12.545	14.040
shark_5	Shark	Association	76	6248	758	8.243	7.414
shark_6	Shark	Association	75	4471	626	7.142	7.178
shark_7	Shark	Association	74	2731	489	5.585	4.015

Table 3: Model parameters and disease results for all networks

Data set	ϕ	$\epsilon \times 10^3$	f_ϕ	R_0 (Pre.)	$\gamma \times 10^4$	R_0 (Sim.)	Error
conference_0	0.631	0.356	0.221	1.254	0.206	1.173	0.187
conference_1	0.521	0.150	0.016	1.234	0.204	1.141	0.250
conference_2	0.288	0.005	0.424	1.054	0.164	0.880	0.222
hospital_0	0.614	0.878	0.769	1.223	0.441	1.144	0.130
hospital_1	0.592	0.825	0.845	1.239	0.635	1.129	0.168
hospital_2	0.681	1.244	0.803	1.279	0.485	1.260	0.128
hospital_3	0.584	0.759	0.713	1.213	0.430	1.127	0.148
school_0	0.620	0.070	1.245	1.378	0.784	1.296	0.139
school_1	0.630	0.076	1.148	1.368	0.792	1.294	0.151
highschool_0	0.422	0.003	0.830	1.175	0.456	0.993	0.233
highschool_1	0.419	0.002	0.685	1.167	0.411	0.981	0.231
highschool_2	0.369	0.001	0.765	1.129	0.330	0.949	0.224
highschool_3	0.367	0.001	0.669	1.130	0.328	0.947	0.237
highschool_4	0.291	0.000	0.345	1.073	0.135	0.829	0.254
ants_1_high	0.962	2.030	0.450	1.556	1.164	1.501	0.110
ants_2_high	0.936	2.105	0.427	1.506	0.812	1.324	0.183
ants_3_high	1.107	2.752	0.655	1.615	1.540	1.598	0.125
ants_1_low	1.116	3.075	0.782	1.606	1.503	1.603	0.088
ants_2_low	0.972	1.926	0.626	1.561	1.202	1.489	0.108
ants_3_low	0.956	1.763	0.548	1.525	1.291	1.498	0.116
Blonder_ants_1_1	0.763	0.794	0.806	1.461	35.825	1.545	0.150
Blonder_ants_1_2	0.761	1.065	0.763	1.450	34.166	1.547	0.144
Blonder_ants_2_1	1.106	2.970	0.825	1.631	23.091	1.707	0.103
Blonder_ants_2_2	0.760	1.124	0.810	1.416	37.926	1.548	0.175
Blonder_ants_6_1	1.190	8.320	1.105	1.604	25.555	1.726	0.145
Blonder_ants_6_2	1.006	6.407	0.751	1.516	16.115	1.619	0.148
bats_0	0.427	3.105	0.640	1.228	6.380	1.083	0.339
voles_BHP	0.666	0.141	0.368	1.412	132.051	1.375	0.169
voles_KCS	0.719	0.206	0.632	1.459	186.728	1.509	0.147
voles_PLJ	0.743	0.183	0.448	1.476	175.470	1.517	0.145
voles_ROB	0.765	0.982	0.243	1.454	95.155	1.290	0.182
parakeet_1_1	1.738	22.574	0.738	1.671	-	-	-
parakeet_1_2	1.437	18.361	0.061	1.619	-	-	-
parakeet_1_3	0.869	8.568	0.268	1.446	-	-	-
parakeet_1_4	1.717	22.304	1.151	1.675	-	-	-
parakeet_2_1	1.148	15.418	0.878	1.532	-	-	-
parakeet_2_2	1.374	19.537	0.957	1.593	-	-	-
parakeet_2_3	1.358	19.266	1.032	1.588	-	-	-
parakeet_2_4	1.157	15.592	1.204	1.551	-	-	-
Bison_dominance	1.058	9.130	0.884	1.546	-	-	-
Sheep_dominance	1.349	12.166	0.940	1.649	-	-	-
cattle_dominance	1.354	12.230	0.946	1.647	-	-	-
macaque_dominance	1.883	7.795	1.442	1.860	-	-	-
Macaques_Massen	0.523	1.573	1.913	1.320	-	-	-
Macaques_Sade	0.350	1.917	0.912	1.208	-	-	-
stumptailed_macaque	0.602	4.776	0.958	1.344	-	-	-
Howler_monkeys	0.538	4.605	1.009	1.321	-	-	-
Kangaroo_proximity	0.507	4.030	1.247	1.245	-	-	-
Swallow_proximity	1.048	15.619	1.578	1.505	-	-	-
shark_0	0.683	1.773	0.707	1.389	-	-	-
shark_1	0.577	0.454	1.080	1.356	-	-	-
shark_2	0.607	0.366	1.153	1.370	-	-	-
shark_3	0.604	0.390	2.009	1.378	-	-	-
shark_4	0.565	0.256	2.442	1.333	-	-	-
shark_5	0.573	0.332	1.929	1.338	-	-	-
shark_6	0.590	0.387	1.727	1.366	-	-	-
shark_7	0.542	0.274	1.161	1.330	-	-	-