




Testing the hypothesis of preferential attachment in social network formation

Thomas House^{1,2*} , Jonathan M Read³, Leon Danon⁴ and Matthew J Keeling²

*Correspondence:

thomas.house@manchester.ac.uk

¹School of Mathematics, University of Manchester, Oxford Road, Manchester, M13 9PL, UK

²Warwick Infectious Disease Epidemiology Research (WIDER), University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK
Full list of author information is available at the end of the article

Abstract

The hypothesis of preferential attachment (PA) - whereby better connected individuals make more connections - is hotly debated, particularly in the context of epidemiological networks. The simplest models of PA, for example, are incompatible with the eradication of any disease through population-level control measures such as random vaccination. Typically, evidence has been sought for the presence or absence of preferential attachment via asymptotic power-law behaviour. Here, we present a general statistical method to test directly for evidence of PA in count data and apply this to data for contacts relevant to the spread of respiratory diseases. We find that while standard methods for model selection prefer a form of PA, careful analysis of the best fitting PA models allows for a level of contact heterogeneity that in fact allows control of respiratory diseases. Our approach is based on a flexible but numerically cheap likelihood-based model that could in principle be applied to other integer data where the hypothesis of PA is of interest.

Keywords: MLE; Phase-type distribution; model selection; spectral methods

1 Introduction

1.1 Contact heterogeneity in infectious disease epidemiology

Infectious pathogens that spread via contact between people are a major cause of human disease, driving attempts to understand their epidemiology [1]. Much theoretical work on infectious disease dynamics has been focused on the role of heterogeneity in the human population [2], which is often conceptualised as a network of epidemiologically relevant contacts [3–5].

Perhaps the most important quantity in any infectious disease outbreak is the basic reproductive ratio, R_0 , which is defined verbally as the expected number of secondary cases generated by an average primary case early in the epidemic. An epidemic is possible exactly when $R_0 > 1$, and typically the efforts required to control such an outbreak grow monotonically with R_0 [1, 2]. In the simplified scenario where each individual picks an integer number of contacts K from the same *degree distribution*, but transmission is otherwise homogeneous,

$$R_0 \propto \mathbf{E}[K^2]. \quad (1)$$

This dependence of the basic reproductive ratio on the second moment of the degree distribution has been a ‘textbook’ result for some time [6], however work by Pastor-Satorras and Vespignani [7] and May and Lloyd [8] raised the question of what might happen for large, or asymptotically divergent, second moments. Such questions can be posed and answered at different levels of mathematical rigour [9] however in the context of epidemiology it is clear that a highly variable degree distribution is associated with the epidemiologically unrealistic scenario that even the most weakly transmissible pathogen can cause an epidemic, and as a corollary that control of any infectious disease through untargeted vaccination would be impossible.

1.2 Data

Of course, whether such a theoretical possibility matters for the study of infectious diseases depends on the actual variance in degree for epidemiologically relevant contacts. While 20th century models of infectious disease were often based on strong a priori assumptions about mixing patterns [1], various methods for measurement of contact patterns now exist and were reviewed by Read et al. [10]. As well as direct measurement of individuals through surveys [11] it is possible to improve coverage through snowball and respondent-driven sampling [12, 13], to make use of the extremely large datasets produced by electronic sensors [14, 15], and also to combine aggregate data [16, 17].

Previous empirical studies have seen evidence that for direct (e.g. [16, 18]) and sexual (e.g. [19, 20]) contacts, an approximate power-law relationship may hold such that for large k , a randomly selected node obeys

$$\Pr(\text{node has } k \text{ links}) \approx k^{-\gamma}. \quad (2)$$

As is the case for almost all biological data, there is much more complexity in the data than such a simple parametric relationships as (2) would imply. For example, Leigh Brown et al. [20] found that while a power-law was a better functional form than the negative binomial for sexual contacts, the richer Waring distribution was preferable to either. What is hard to dispute, however, is that as better quality data on epidemiologically relevant contacts is obtained the evidence consistently points to a very high level of variance; for example, Read et al. [21] were able to validate the high numbers of contacts reported by some study participants through direct (rather than postal) surveying.

These empirical observations of high heterogeneity in contact number, together with theoretical results about R_0 , present a paradox for infectious disease epidemiology: is the extreme heterogeneity in observed contact patterns indicative of PA and does that imply that $R_0 > 1$ for almost any finite level of person-to-person transmissibility meaning that our theoretical understanding of infectious disease epidemiology is somehow severely lacking?

1.3 Preferential attachment and power laws in empirical data

Recent years have seen a debate about the level of heterogeneity that exists in a variety of observed networks. A particularly influential paper by Barabási and Albert [22] considered a model of network formation in which many new nodes are added to a small existing network. These new nodes connect preferentially to nodes that have more links in the existing network, leading to the asymptotic result (2) with $\gamma = 3$. In this way preferential

attachment is intimately linked with, but not always equivalent to, asymptotic power-law behaviour.

Simple power-law relationships have been claimed for numerous real-world systems, and a critical review of these claims by Clauset et al. [23] used maximum-likelihood fitting of distribution tails to power-law distributions to show varying levels of statistical support for claims in the literature. In the context of discrete data, pioneering work by Zipf [24] found power-laws in word frequencies; considering the count of unique words in *Moby Dick* both Newman [25] and Clauset et al. [23] agree that the statistical evidence for Zipf's power-law distribution in this context is strong. On the other hand, the in- and out-degrees of *E. coli* metabolic networks have been claimed to follow a power law [26], but this is disputed by the analyses of Huss and Holme [27] and Clauset et al. [23].

The debate around presence or absence of power laws in real data continues, perhaps most strongly in the context of networks. For example, Barabási [28] writes that preferential attachment is network science's "most profuse concept," and that "the impact of preferential attachment is hard to miss." At the same time, Stumpf and Porter [29] argue that "most reported power laws lack statistical support and mechanistic backing."

1.4 Testing preferential attachment directly

In this work, we attempt to test the hypothesis of preferential attachment in social contact data directly, rather than via asymptotic power law behaviour. We make use of previously collected data on social encounters specifically designed to measure heterogeneity in numbers of contacts amongst the British population, and fit mechanistic models of different complexity to these data. We determine that models with significant levels of preferential attachment have better evidential support from the data than models without.

2 Methods

2.1 Social Contact Survey data

A cross-sectional study was conducted between May 2009 and October 2010, recruiting households and individuals through postal and online questionnaires, supported by a large random-address mailshot and a modest online and media promotion [30, 31]. Questionnaires asked respondents to report on the number of distinct individuals they encountered the previous day: their contacts. Respondents were able to report contacts either as individuals or as members of a group with a reported size. Allowing the reporting of groups of individuals was a deliberate methodological design to permit the easy reporting of large numbers of contacts, to avoid the approach taken by previous studies [11], which imposed a high burden on respondents with large number of contacts, and to ensure the best capture of the right-hand tail of the degree distribution. In general, we expect that such data will become increasingly available due to the epidemiological importance of this tail (e.g. the study of Read et al. [21]).

In total, completed questionnaires were received from 5,388 participants in Great Britain, 3,901 of which were from postal surveys. There was some bias in demographic representation, most notably younger age groups and males were generally under-represented (see Danon et al. [31] for more details). The data is available at <http://wrap.warwick.ac.uk/54273/>.

2.2 Generalised preferential attachment

As noted by Durrett [32], Barabási [28], and Simkin and Roychowdhury [33], the basic idea behind the preferential attachment model is close to the population model of Yule [34]. We consider a Yule-like stochastic process described precisely as follows. In a population of N individuals indexed by i each individual has an integer-valued random variable $K_i(t)$ for its number of contacts. Individual i starts with $K_i(0) = 0$ and makes new contacts over a time period T_i , which is given by a positive real-valued random variable with probability density function $\rho(t)$. The generation of new social contacts is modelled by a continuous-time Markov chain with the following events and rates:

$$K_i \rightarrow K_i + 1 \quad \text{at rate } f_{K_i} := 1 + \tau K_i. \quad (3)$$

We take the preferential attachment hypothesis PA to be stated mathematically as

$$\text{PA} \Leftrightarrow \tau > 0. \quad (4)$$

Writing $p_k(t)$ for the probability that $K_i = k$ at time $t < T_i$, we can use the method of characteristics to derive an expression for the probability generating function of K_i ,

$$g(t, s) = \sum_k p_k(t) s^k = (s - (s - 1)e^{\tau t})^{-1/\tau}, \quad s \in [0, 1]. \quad (5)$$

From this, we can derive expressions for the probability mass function,

$$\begin{aligned} p_k(t) &= \frac{1}{k!} \left. \frac{\partial^k g}{\partial s^k} \right|_{s=0} = \frac{\Gamma(k + \frac{1}{\tau})}{\tau \Gamma(\frac{1}{\tau}) \Gamma(k + 1)} e^{-t} (e^{\tau t} - 1)^k \\ &\rightarrow \kappa_t (e^{\tau t} - 1)^k k^{(1-\tau)/\tau}, \end{aligned} \quad (6)$$

where κ_t is a function of t but not k and the asymptote holds as k becomes large. This is not a simple power-law relationship, and so the asymptotic behaviour of the moments is not determined by the power-law exponent, but rather through the moment generating function $M(t, z) = g(t, e^z)$, $z \in (-\infty, 0]$, such that the r th moment of the degree distribution, conditional on $t < T$, is

$$m_r(t) = \left. \frac{\partial^r M}{\partial z^r} \right|_{z=0}. \quad (7)$$

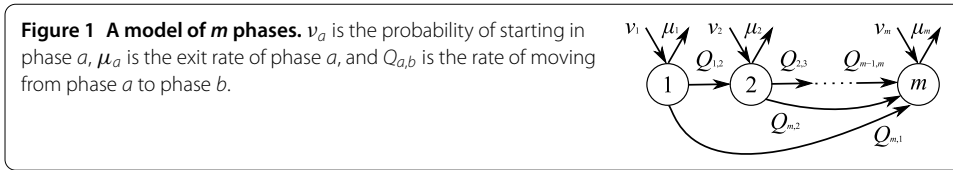
In particular,

$$m_1(t) = \frac{1}{\tau} (e^{\tau t} - 1), \quad m_2(t) = m_1(t) + (\tau + 1)(m_1(t))^2, \quad \dots \quad (8)$$

Then accounting for the randomness of the times, the r th moment of the degree distribution will be

$$\bar{m}_r = \int_{t=0}^{\infty} \rho(t) m_r(t) dt. \quad (9)$$

We will be interested in the empirical evidence for whether such moments converge or diverge, in light of the epidemiological relationship (1).



2.3 Phase-type holding times

The question is then posed as to an appropriate distribution from which to draw the holding times $\{T_i\}$ for the amount of time spent making new contacts on the day for which individuals provide data. In previous work [30] on a related model of contact formation we considered holding times T_i that were log-normally distributed. This provided a good fit to data, but was computationally intensive and lacked a mechanistic interpretation. We therefore consider here a class of distributions for the holding times that is highly flexible, but which has analytic and numerical benefits - the distributions of *phase type* [35]. Phase-type distributions are dense in the space of positive-valued probability distributions [36], meaning that they can be made arbitrarily close to any other distribution. They have a mechanistic interpretation and allow for analytic manipulations that greatly reduce the numerical cost of likelihood evaluation.

The basic idea behind the model is shown in Figure 1. A set of phases is indexed by $a, b = 1, \dots, m$; the probability of starting in phase a is v_a (meaning these parameters must sum to unity); the rate of stopping making new social contacts is μ_a for an individual in phase a ; and the rate of moving from phase a to phase b is $Q_{a,b}$. Note that different rates of making contacts in different phases are not realistically distinguishable from different times spent and so are not included as parameters. The phases have a mechanistic interpretation as the activities that individuals undertake on a given day.

In this model, the probability density function for the holding time is given by the general expression

$$\rho(t) = \boldsymbol{\mu}^\top e^{\mathbf{M}t} \mathbf{v}, \tag{10}$$

where:

$$\begin{aligned} \boldsymbol{\mu} &= (\mu_a); & \mathbf{v} &= (v_a); \\ M_{a,a} &= -M_a; & M_a &= \mu_a + \sum_b Q_{b,a}; & M_{a,b \neq a} &= Q_{b,a}. \end{aligned} \tag{11}$$

From the expressions (5), (7), (8) and (9) above, in particular through inspection of the form of the moment generating function, it is clear that the r th moment of the degree distribution will involve a term like

$$\mathcal{I}_r = \int_{t=0}^{\infty} e^{r\tau t} \rho(t) dt = \boldsymbol{\mu}^\top \int_{t=0}^{\infty} e^{(\mathbf{M}+r\tau\mathbf{I})t} dt \mathbf{v}, \tag{12}$$

where \mathbf{I} is the identity matrix. Let $\mathbf{A} = \mathbf{M} + r\tau\mathbf{I}$; this matrix is triangular and so its eigenvalues are equal to its diagonal elements; in particular the a th eigenvalue of \mathbf{A} is

$\lambda_a = -M_a + r\tau$. If we let \mathbf{R} be a matrix whose a th column is the a th right eigenvector of \mathbf{A} and \mathbf{L} be a matrix whose a th row is the a th left eigenvector of \mathbf{A} then

$$\mathcal{I}_r = \boldsymbol{\mu}^\top \int_{t=0}^{\infty} \mathbf{L}^{-1} \mathbf{L} e^{\mathbf{A}t} \mathbf{R} \mathbf{R}^{-1} dt \mathbf{v} = \boldsymbol{\mu}^\top \mathbf{L}^{-1} \int_{t=0}^{\infty} \mathbf{D} dt \mathbf{R}^{-1} \mathbf{v}, \quad (13)$$

where \mathbf{D} is a diagonal matrix whose a th diagonal element is $e^{\lambda_a t}$. The integral \mathcal{I}_r therefore converges exactly when $\forall a, \lambda_a < 0$, which implies that the condition for divergence of the r th moment is

$$\bar{m}_r \text{ diverges} \Leftrightarrow \exists a \text{ such that } \tau > M_a/r. \quad (14)$$

In general, however, combination of (10) and (6) is not the most numerically efficient method for calculation of the overall probability mass function for final number of contacts $K_i(T_i)$ and a different approach is needed.

2.4 Numerically efficient model solution

The model as described above can be solved in a numerically efficient manner using the spectral methods for continuous-time Markov chains developed by Bailey [37]. We consider the limit as the population size $N \rightarrow \infty$ and write down ordinary differential equations (ODEs) for the proportion of the population in phase a and with k social contacts at time t , $p_{a,k}(t)$. These ODEs take the form

$$\frac{d}{dt} p_{a,k} = - \left(f_k + \mu_a + \sum_{b>a} Q_{a,b} \right) p_{a,k} + f_{k-1} p_{a,k-1} + \sum_{b<a} Q_{b,a} p_{b,k}, \quad (15)$$

where f_k is the rate at which individuals with k social contacts make new contacts, given in (3). We are then interested in d_k , the probability mass function for a randomly selected individual having made k social contacts by the end of the process. A series of manipulations directly analogous to those of Bailey [37] shows that

$$d_k = \lim_{s \downarrow 0} \sum_a \mu_a \int_0^{\infty} e^{-st} p_{a,k}(t) dt =: \sum_a \mu_a A_{a,k}. \quad (16)$$

Applying Laplace transformation to (15) subject to the initial condition $p_{a,k}(0) = \nu_a \delta_{k,0}$ and taking the frequency-space limit $s \downarrow 0$ then leads to a set of linear equations for d_k that are triangular and so can be evaluated directly without numerically costly matrix inversion:

$$\nu_a \delta_{k,0} = - \left(f_k + \mu_a + \sum_{b>a} Q_{a,b} \right) A_{a,k} + f_{k-1} A_{a,k-1} + \sum_{b<a} Q_{b,a} A_{b,k}. \quad (17)$$

These equations are at the root of the numerical efficiency of our model. Note that we use Laplace transformation mainly for technical reasons and our results could be obtained by directly integrating (15) if one were not concerned by all quantities being rigorously defined.

2.5 Model likelihood, fitting and selection

We assume a vector of data $\mathbf{y} = (y_k)$, where y_k is the number of individuals reporting k social contacts when surveyed. A model \mathcal{M} is therefore specified by a number of phases m and the presence or absence of PA, meaning the general parameters are $\boldsymbol{\theta} = (\tau, \nu_a, \mu_a, Q_{a,b})$, with τ present only if there is PA. The number n of individuals sampled from the British population N is

$$n = \sum_k y_k = 5388 \ll N \gtrsim 6 \times 10^7, \quad (18)$$

and so it is appropriate to assume that each individual picks a number of contacts independently from the distribution with pmf given by d_k as in (16). Accounting for the censoring of zero contacts in the real data, we define

$$\tilde{d}_0 = 0, \quad \tilde{d}_{k>0} = \frac{d_k}{1 - d_0}, \quad (19)$$

meaning that the overall likelihood function is then given by

$$L(\mathbf{y}|\boldsymbol{\theta}) = \frac{n!}{\prod_l y_l!} \prod_k (\tilde{d}_k(\boldsymbol{\theta}))^{y_k}. \quad (20)$$

Note that the combinatorial factors do not depend on the parameters, and so need not be calculated during model fitting.

We consider the use of the likelihood function (20) using standard statistical methodology. Numerical maximum likelihood estimation was performed using simulated annealing run from multiple starting points to ensure the global optimum was obtained. Model selection was performed using AIC [38] and BIC [39], as well as likelihood ratio tests [40] on pairs of models where this test was informative. This was done since each approach involves different trade-offs between model fit and complexity, and to check that our conclusions about PA are not overly sensitive to the precise method used. Uncertainty in model parameters was quantified using confidence intervals obtained through bootstrapping the data, and uncertainty in model outputs such as the predicted degree distribution was quantified using a parametric bootstrap.

3 Results and discussion

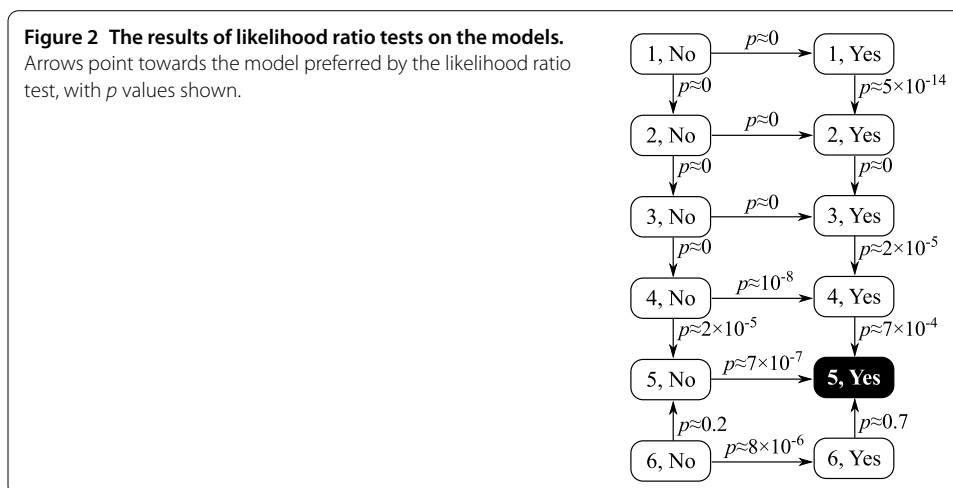
Table 1 shows the models we fitted, their number of parameters, AIC/BIC relative to the minimum, and the first moment that diverges according to (14). Figure 2 shows the results of performing likelihood ratio tests. These show that AIC prefers a 5-phase model with PA as do likelihood ratio tests for any significance level between 0.07% and 20%. BIC penalises complex models more severely and therefore selects a 3-phase model with PA.

Therefore, regardless of the number of phases selected by different approaches to model selection, we see that the models with PA are preferred over models without. Figure 3 shows the predictions of the models preferred by different selection criteria, as well as the models with the same number of phases but no PA, against real data. We see in the left column of plots that for the 5-phase models, the main difference is in the tail of the distribution as we would expect. In the 3-phase models shown in the right column of plots, the model without PA also smooths over features in the bulk of the distribution compared to the model with PA.

Table 1 Comparison of models with different numbers of phases, with and without preferential attachment (PA), together with: number of parameters; differences in AIC and BIC values compared to the overall minimum; and the lowest divergent moment for models with PA

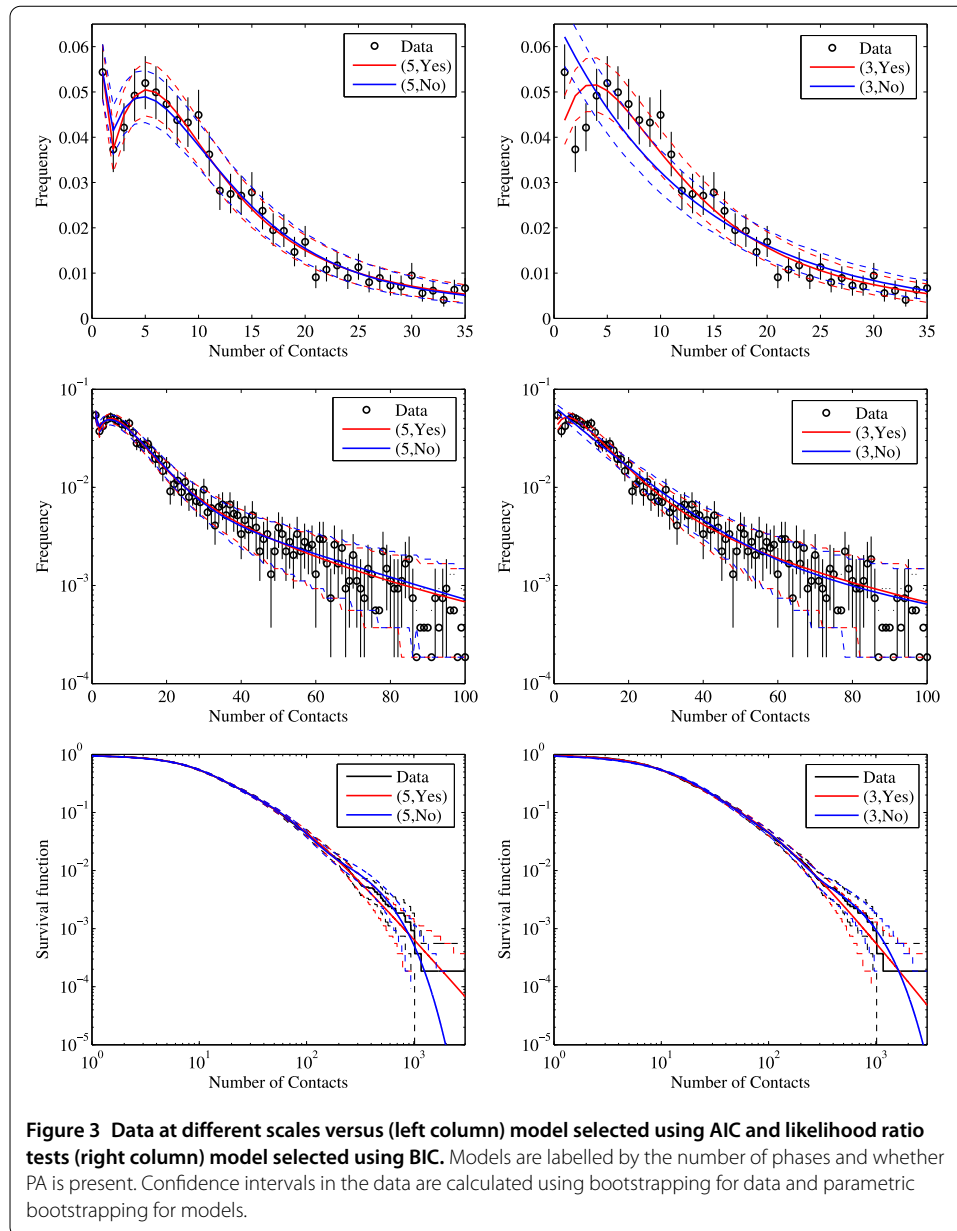
(Phases, PA)	No. Params	ΔAIC	ΔBIC	Diverge
(1, No)	1	2.2×10^3	2.1×10^3	–
(2, No)	4	2.1×10^2	1.5×10^2	–
(3, No)	8	1.2×10^2	83	–
(4, No)	13	42	38	–
(5, No)	19	23	58	–
(6, No)	26	27	1.1×10^2	–
(1, Yes)	2	1.9×10^2	1.1×10^2	3
(2, Yes)	5	1.3×10^2	72	4
(3, Yes)	9	31	[0]	3
(4, Yes)	14	11	14	3
(5, Yes)	20	[0]	42	3
(6, Yes)	27	9	97	3

Preferred models are shown using square brackets and bold type.



For the 3-phase model with PA, $\tau = 0.018$ [0.012, 0.026]; and if we set $\tau = 0$ but leave the other parameters at their fitted values, then the total number of contacts per person is reduced to 64% of its original value. For the 5-phase model with PA $\tau = 0.026$ [0.019, 0.036]; and if we set $\tau = 0$ but leave the other parameters at their fitted values, then the total number of contacts per person is reduced to 58% of its original value. This shows that in both of these models, we can attribute a substantial fraction of the contacts to PA.

We also calculate that the second moment does not diverge in any of the fitted models, which helps to resolve the epidemiological paradox that we introduced at the start of this paper. PA is empirically supported, and is also mechanistically plausible since existing social contacts give more opportunities for future social contact. Combined with a sufficiently detailed phase-based mechanistic model of the contexts in which social contacts are made, however, PA does not imply a divergent second moment for the distribution of contacts relevant for the spread of directly transmitted infections. This means that our understanding of how basic epidemiological quantities like the basic reproductive ratio, R_0 , are related to contact networks does not need to be revised in the light of empirical evidence.



As a final observation, we believe that as computational resources for fitting models to data improve, it will in general be easier to test the hypothesis of PA directly in all kinds of data, rather than looking for asymptotic power laws.

Appendix A: Code

We provide the function `phase_ll.c` below as a Mex file for use within MATLAB; however, the C syntax is standard and modification to run within R, Python or C should be straightforward. Once compiled, use within MATLAB is:

```
>> [LL, P] = phase_ll(params, M, kh, K);
```

where `params` is a vector of the model parameters `[nu(:), mu(:), Q(:), tau]`, `M` is the number of phases, `kh` is a vector of count frequencies from `0 : K`, `K` is the maximum

count, LL is the log-likelihood given the parameters and P is the probability vector for the observed counts given the parameters.

A.1 The function `phase_ll.c`

```
#include <mex.h>
#include <math.h>

void mexFunction (int nlhs, mxArray *plhs[],
                  int nrhs, const mxArray *prhs[])
{
    /* Input variables */
    double *params, *kh, *Mmex, *Kmex;
    size_t lp;
    int M, K;

    /* Get inputs in correct form */
    params = mxGetPr(prhs[0]);
    lp = mxGetN(prhs[0]);
    Mmex = mxGetPr(prhs[1]);
    kh = mxGetPr(prhs[2]);
    Kmex = mxGetPr(prhs[3]);
    M = (int) Mmex[0];
    K = (int) Kmex[0];

    /* Working variables */
    double nu[M], mu[M], Q[M][M], s[M];
    double f[K];
    double tau, S;
    double A[K][M];
    int fi, fj, k, m;

    /* Output variables */
    double *P, *LL;

    /* Organise outputs into correct form */
    plhs[0] = mxCreateDoubleMatrix(1,1,mxREAL);
    LL = mxGetPr(plhs[0]);
    plhs[1] = mxCreateDoubleMatrix(K,1,mxREAL);
    P = mxGetPr(plhs[1]);

    nu[0] = 1.0;
    for(fi=1; fi<M; fi++) {
        nu[fi] = params[fi-1];
        nu[0] -= params[fi-1];
    }
    for(fi=M; fi<(2*M); fi++) {
        mu[fi-M] = params[fi-1];
    }
    m = (2*M)-1;
    for(fi=0; fi<M; fi++) {
        s[fi] = 0;
        for(fj=(fi+1); fj<M; fj++) {
```

```

        Q[fi][fj] = params[m];
        s[fi] += params[m];
        m=m+1;
    }
}
if (lp == (m+1))
    tau = params[m];
else
    tau = 0;

f[0] = 1;
for(k=1; k<(K-1); k++)
    f[k] = 1+(tau*k);
f[K-1] = 0;

A[0][0] = (nu[0])/(f[0]+mu[0]+s[0]);
for(k=1; k<K; k++) {
    A[k][0] = ( f[k-1]/(f[k]+mu[0]+s[0]) ) * A[k-1][0];
}
for(m=1; m<M; m++) {
    S=0.0;
    for(fi=0; fi<m; fi++) {
        S += Q[fi][m] * A[0][fi];
    }
    A[0][m] = (nu[m]+S)/(f[0]+mu[m]+s[m]);
    for(k=1; k<K; k++) {
        S=0.0;
        for(fi=0; fi<m; fi++) {
            S += Q[fi][m] * A[k][fi];
        }
        A[k][m] = ((f[k-1] * A[k-1][m]) + S) / (f[k] + mu[m] + s[m]);
    }
}

P[0] = 0.0;
for(m=0; m<M; m++) {
    P[0] += A[0][m] * mu[m];
}

for(k=1; k<K; k++) {
    P[k] = 0.0;
    for(m=0; m<M; m++) {
        P[k] += A[k][m] * mu[m];
    }
    P[k] /= (1-P[0]);
}
P[0] = 0.0;

LL[0] = 0.0;
for(k=1; k<(K-1); k++) {
    LL[0] += kh[k] * log(P[k]);
}
}

```

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors: collected the data, designed the study. TH: performed the analysis.

Author details

¹School of Mathematics, University of Manchester, Oxford Road, Manchester, M13 9PL, UK. ²Warwick Infectious Disease Epidemiology Research (WIDER), University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK. ³CHICAS, Faculty of Health and Medicine, Lancaster University, Lancaster, Lancashire, LA1 4YG, UK. ⁴School of Social and Community Medicine, University of Bristol, Oakfield Grove, Clifton, BS8 2BN, UK.

Acknowledgements

The Social Contact Survey was funded by the Medical Research Council, grant number G0701256. TH and MJK are supported by the Engineering and Physical Sciences Research Council. JMR and MJK are supported by the Economic and Social Research Council, grant ES/K004255/1. LD is supported by the Leverhulme Trust.

Received: 2 July 2015 Accepted: 28 September 2015 Published online: 09 October 2015

References

- Anderson RM, May RM (1991) *Infectious diseases of humans*. Oxford University Press, Oxford
- Diekmann O, Heesterbeek JAP, Britton T (2012) *Mathematical tools for understanding infectious disease dynamics*. Princeton University Press, Princeton
- Bansal S, Grenfell BT, Meyers LA (2007) When individual behaviour matters: homogeneous and network models in epidemiology. *J R Soc Interface* 4(16):879-891
- Danon L, Ford AP, House T, Jewell CP, Keeling MJ, Roberts GO, Ross JV, Vernon MC (2011) Networks and the epidemiology of infectious disease. *Interdiscip Perspect Infect Dis* 2011:284909
- Pellis L, Ball F, Bansal S, Eames K, House T, Isham V, Trapman P (2014) Eight challenges for network epidemic models. *Epidemics* 10:58-62. doi:10.1016/j.epidem.2014.07.003
- Diekmann O, Heesterbeek JAP (2000) *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Wiley, New York
- Pastor-Satorras R, Vespignani A (2001) Epidemic dynamics and endemic states in complex networks. *Phys Rev E* 63:066117
- May RM, Lloyd AL (2001) Infection dynamics on scale-free networks. *Phys Rev E* 64:066112
- Durrett R (2010) Some features of the spread of epidemics and information on a random graph. *Proc Natl Acad Sci USA* 107(10):4491-4498
- Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DAT (2012) Close encounters of the infectious kind: methods to measure social mixing behaviour. *Epidemiol Infect* 140(12):2117-2130
- Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, Massari M, Tomba GS, Wallinga J, Heijne J, Sadkowska-Todys M, Rosinska M, Edmunds WJ (2008) Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med* 5(3):381-391
- Goodman LA (1961) Snowball sampling. *Ann Math Stat* 32:148-170
- Heckathorn DD (1997) Respondent-driven sampling: a new approach to the study of hidden populations. *Soc Probl* 44:174-199
- Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, Jones JH (2010) A high-resolution human contact network for infectious disease transmission. *Proc Natl Acad Sci USA* 107(51):22020-22025
- Isella L, Stehlé J, Barrat A, Cattuto C, Pinton J, Van den Broeck W (2011) What's in a crowd? Analysis of face-to-face behavioral networks. *J Theor Biol* 271(1):166-180
- Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, Toroczkai Z, Wang N (2004) Modelling disease outbreaks in realistic urban social networks. *Nature* 429(6988):180-184
- Eubank S, Barrett C, Beckman R, Bisset K, Durbeck L, Kuhlman C, Lewis B, Marathe A, Marathe M, Stretz P (2010) Detail in network models of epidemiology: are we there yet? *Journal of Biological Dynamics* 4(5):446-455
- Fournet J, Barrat A (2014) Contact patterns among high school students. *PLoS ONE* 9(9):e107878
- Schneeberger A, Mercer CH, Gregson SAJ, Ferguson NM, Nyamukapa CA, Anderson RM, Johnson AM, Garnett GP (2004) Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in Britain and Zimbabwe. *Sex Transm Dis* 31(6):380-387
- Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT (2011) Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J Infect Dis* 204(9):1463-1469
- Read JM, Lessler J, Riley S, Wang S, Tan LJ, Kwok KO, Guan Y, Jiang CQ, Cummings DAT (2014) Social mixing patterns in rural and urban areas of southern China. *Proc R Soc B* 281(1785):20140268
- Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509-512
- Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. *SIAM Rev* 51(4):661-703
- Zipf GK (1949) *Human behavior and the principle of least effort: an introduction to human ecology*. Addison-Wesley, Reading
- Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. *Contemp Phys* 46(5):323-351
- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. *Nature* 407(6804):651-654
- Huss M, Holme P (2007) Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. *IET Syst Biol* 1(5):280-285
- Barabási AL (2012) Network science: luck or reason. *Nature* 489(7417):507-508
- Stumpf MPH, Porter MA (2012) Critical truths about power laws. *Science* 335(6069):665-666
- Danon L, House T, Read JM, Keeling MJ (2012) Social encounter networks: collective properties and disease transmission. *J R Soc Interface* 9(76):2826-2833

31. Danon L, Read JM, House T, Vernon MC, Keeling MJ (2013) Social encounter networks: characterizing Great Britain. *Proc R Soc B* 280(1765):20131037
32. Durrett R (2007) *Random graph dynamics*. Cambridge University Press, Cambridge
33. Simkin MV, Roychowdhury VP (2011) Re-inventing Willis. *Phys Rep* 502(1):1-35
34. Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philos Trans R Soc Lond B, Contain Pap Biol Character* 213:21-87
35. Neuts MF (1981) *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Johns Hopkins University Press, Baltimore
36. Neuts MF (1975) Probability distributions of phase type. In: *Liber amicorum Professor emeritus Dr. H. Florin*. Katholieke Universiteit Leuven, Departement Wiskunde, Leuven, pp 173-206
37. Bailey NTJ (1957) *The mathematical theory of epidemics*. Griffin, London
38. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Autom Control* 19(6):716-723
39. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461-464
40. Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc Lond A, Math Phys Eng Sci* 231:289-337

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Immediate publication on acceptance
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ springeropen.com
