Electronic supplementary material to *Weighting for sex acts to understand the spread of STI on networks*

Mathieu Moslonka-Lefebvre[a,b,*], Sebastian Bonhoeffer[a], Samuel Alizon[c]

[a] *Institut for Integrative Biology, ETH, Universitätstrasse 16, 8092 Zürich, Switzerland*
[b] *Present address: INRA, UR 341 Mathématiques et Informatique Appliquées, 78350 Jouy-en-Josas, France*
[c] *Laboratoire MIVEGEC (UMR CNRS 5290, IRD 224, UM1, UM2), 911 avenue Agropolis, 34394 Montpellier Cedex 5, France*

Contents

*Corresponding author, email: mathieu.moslonka-lefebvre@jouy.inra.fr

S1. Supplementary Methods

S1.1. About the Networks

S1.1.1. Introduction to networks

In this subsection, we only present some basic elements of network theory. For further reading, the reader should refer to more detailed reviews [e.g. 1, 2]. Concerning the applications of network theory to epidemiology, see Newman [3] or Keeling and Eames [4].

Network theory has been used in recent years to help understand and predict the behaviour of a range of natural, social and technological systems [1, 2]. A network of N interacting units (e.g. individuals, species, nurseries, farms, hospitals or airports) is composed of N nodes, or vertices, interacting with each other by edges, or links, the connections being here purely relational and not necessarily related to the Euclidean distance between two entities [5]. In some networks, edges and nodes can be assigned weights.

When modelling spread of pathogens on networks, we consider that nodes are individuals and links are interactions among them (e.g. sexual intercourse, handshake, etc.). Our work is applied to sexually transmitted infections (STIs) in general but we focus on the case of HIV. Moreover, we only study homosexual networks which are unipartite networks (either men-men or women-women interactions only), and not, heterosexual and bisexual networks which are bipartite networks [6]. Our model can nevertheless be extended to any type of network, including heterosexual networks.

We also assume that the network is static, i.e. that the interactions between individuals do not vary over time. However, our model can be extended to consider dynamical networks such as concurrent partnerships

networks [7].

Mathematically, a static network can be represented by an $N \times N$ adjacency matrix $\mathcal{A}$ whose elements $\mathcal{A}(i,j)$ are equal to 1 if there is a link connecting node i to node j and 0 otherwise. Here, a node cannot be connected to itself (self-loops are not biologically relevant) so $\mathcal{A}(i,i) = 0$. We also assume that networks are undirected, i.e. that $\mathcal{A}(i,j) = \mathcal{A}(j,i)$. This might not be the case for human immunodeficiency virus (HIV) as transmission rates are known to differ between partners (e.g. men and women in heterosexual couples). Using directed networks would increase the realism [8] but would also make the model much more difficult to analyse. Note that the weighting model presented below can be straightforwardly applied to directed networks.

Four main types of networks can be defined from the topology of a network, i.e. depending on the repartition of the numbers 0 and 1 in the adjacency matrix [1, 5, 9]: local networks, where edges only link adjacent nodes, random networks, where nodes are connected randomly, small-world networks, which are topologically local networks with some nodes rewired randomly, thus generating heterogeneous networks in which a few nodes are highly connected while the others nodes have few connections. Our focus here is on heterogeneous networks. We focus in particular on networks following a power law or a negative binomial distribution but we also show that a similar approach can be applied to homogeneous small-world networks.

S1.1.2. Terminology

Degree. For any node i part of a network, the degree k_i of i is the total number of edges from node i to all other nodes. We define the degree vector

$\mathbf{k}$ such that $\mathbf{k}[i] = k_i$.

Scaling exponents of power law (PL) networks. In **power law networks**, the distribution of the degree k, is assumed to follow a power-law [10], which can be expressed according to Goldstein et al. [11] as

$$p(k, \gamma) = \frac{k^{-\gamma}}{\zeta(\gamma)}$$

where γ is the scaling exponent of the power law ($\gamma > 1$) and ζ is the Riemann zeta function defined by

$$\zeta(\gamma) = \sum_{k=1}^{\infty} k^{-\gamma}$$

Preferential attachment is a well-know mechanism leading to power-laws distributions of degrees in power law networks when the number of nodes added is high [12, 13].

S1.1.3. The empirical GC sexual-contact network

This network (denoted GC in the following) was derived by Potterat et al. [14, 15]. Data were collected by contact-tracing of partners of individuals who have been infected with Gonorrhoea between January 1 and June 30, 1981 in Colorado Springs, Colorado. The age range (in years) of the respondents in this study was [15 - 34] [14]. 1729 individuals were included in the data and connected through sexual connection; 259 appeared only as cases, 993 only as contacts, and 447 as both cases and contacts [14]. 451 men and 285 women were infected with gonorrhoea. Of 451 men diagnosed with gonorrhoea, 69 named (or were named by) homosexual partners and 382 did not (or were not). The former are classified as homosexual and the latter as heterosexual [15].

The GC 1981 network used for analysis is composed of 1726 individuals and 1430 links. The largest connected component was composed of 381 individuals and 381 links. We only focus on the largest connected component.

We used a model comparison based on the Akaike Information Criterion (AIC) to determine the type of distribution that best fitted the node degree distribution of the largest connected component of the GC network. The maximum likelihood estimation for each distribution was perfomed using the `mle2` function from the `bbmle` R package [16]. Results are shown in Table 1.

Table 1: Fitting the best distribution for the node degree distribution of the GC network a and b are parameters fitted using the `bbmle` package in R. Confidence intervals are shown in brackets. k is the node degree.

Distribution	Formula	Best fit	AIC
Poisson	$a^n \frac{e^{-a}}{k!}$	$a = 1.02$ $[0.67, 1.46]$	-32.5
Negative Binomial	$b^k(1-b)^a \binom{a+k-1}{k-1}$, with $\frac{ab}{1-b} = 2$	$b = 0.32$ $[-0.52, 0.73]$	-26.3
Yule	$\frac{a\Gamma(a+1)}{(a+k)^{1+a}}$	$a = 1.99$ $[1.68, 2.05]$	-34.8
Pareto	$\frac{a}{k^{a+1}}$	$a = 0.60$ $[0.57, 0.63]$	-75.2
Truncated Negative Binomial	$\frac{(1-b)^{-a}}{(1-b)^{-a}-1} b^k(1-b)^a \binom{a+k-1}{k-1}$	$a = 0.00061$, $b = 0.67$	-82.4
Zeta Power Law	$\frac{k^{-a}}{\zeta(a)}$	$a = 1.98$ $[1.94, 2.03]$	-90.9

The best fit was obtained with the zeta power law distribution and the corresponding scaling exponent was $\gamma = 1.98$, with a confidence interval 1.94

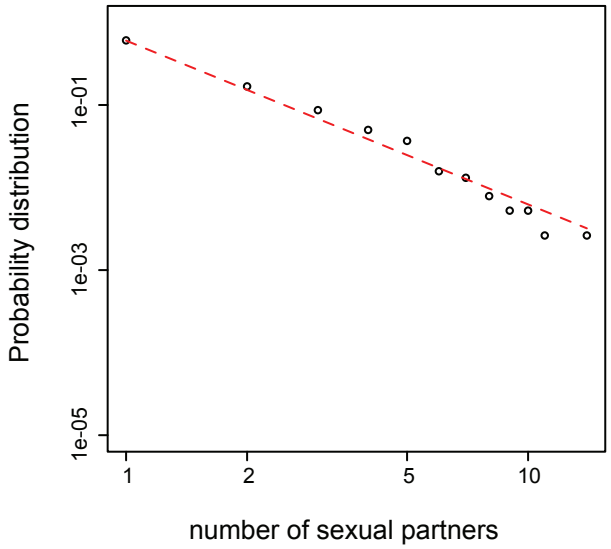

Figure S1: Node degree distribution for the GC network.
The red dashed line indicates the power law distribution.

– 2.03 (see also Table 2). This is illustrated by Figure S1. Note that the AIC difference between the Negative Binomial distribution truncated in 0 and the PL distribution is 8.5 and a difference between models is often considered to be significant if the AIC differ by 10. However, the expectancy of the distribution found in the Truncated NB is of 1.83, which is slightly different from the average network degree (which is 2).

Contact tracing studies have known biases described in [17]. One of these biases is that it is often difficult to identify all the partners of an infected individual. A second bias is that these networks tend to isolate sub-networks where disease spread is easier. However, this does not invalidate the results for such sub-groups of the population because, as shown by Newman

[18], large sexual-contact networks can be seen as many small and dense networks (or 'modules') that are weakly interconnected. Contact tracing studies provide us with a description of one of these modules.

Our goal in this study is to compare networks weighted using different biological assumptions. The GC network offers the great advantage to be drawn from empirical data. However, we want to show that our method is more general and can be applied to heterogeneous networks in general without affecting the results qualitatively. Moreover, the GC network includes homosexuals and heterosexuals, whereas in our approach we assume for simplicity that we are working on a homosexual network (this way we do not need to use a bipartite network). For these two reasons, we also randomly generated heterogeneous networks, as further described below. For these networks, we chose an average degree of 4 because the estimates for the average number of partners found in the literature vary from 2 to 8 partners on average (see below). Finally, in order to compare the results from the GC network and the theoretical networks we set the average number of sex act per week per individual constant and equal to 4 (i.e. to the degree in the theoretical networks).

Table 2: Key properties of the GC network

N is the number of nodes, $\overline{k}$ is the mean degree, $\Lambda_{\mathcal{A}}$ is the dominant eigenvalue of the adjacency matrix $\mathcal{A}$ and γ is power law of coefficient.

Name	N	Edges	$\overline{k}$	$\Lambda_{\mathcal{A}}$	γ
GC1981	381	381	2	3.88	1.98

S1.1.4. Randomly generated power law (PL) networks

In order to better assess the robustness of our results, we generated random power law networks. We chose to study this typical structure because some sexual contact networks have been reported to be power law (see the main text). However, we also used other distributions for heterogenous networks, such as the negative binomial distribution (see below).

The theoretical power law networks we used were built using the Barabasi-Albert algorithm [10], where a basic network called a seed is expanded recursively into a network of greater size with preferential attachment (on the whole, new nodes are added to already highly connected nodes; more precisely, new nodes and links are added proportionally to existing nodes' degrees). In this case, the γ coefficient is 3.

We used an average degree of 4, which is the mean number of partners over the last 5 years for heterosexual men in Britain in 2000 [19] or the lifetime median number of partners for heterosexual women in Englan in 2000 [19] and in France in 2006 [20]. Note that this value can be higher for other communities (for instance, men having sex with men communities tend to have higher number of partners).

S1.1.5. Threshold properties of power law networks

As discussed below, the dominant eigenvalue of the binary adjacency matrix is an indicator of disease spread on networks. For canonical Barabasi-Albert power law networks [10], the dominant eigenvalue $\Lambda_{\mathcal{A}}$ of the binary adjacency matrix $\mathcal{A}$ increases with number of nodes N such that

$$\Lambda_{\mathcal{A}} \approx \sqrt{m}\, N^{\frac{1}{4}}$$

thus leading to an increase in disease speed of spread with N [21]. However, there is a controversy about the relevance of this result to real epidemiological networks [22].

S1.1.6. Randomly generated networks with a negative-Binomial (NB) node degree distribution

Networks reflecting human sexual contacts in a population over several years are often argued to exhibit heterogeneous structures which are e.g. corresponding to 'negative-binomial' (NB) graphs (see the main text).

The negative binomial law $NB(a, b)$ is given by [23]:

$$Pr[k] \;=\; \binom{a+k-1}{k-1} b^k (1-b)^a \tag{S1}$$

where k is an integer for the node degree. The law has a mean $\mu = ab/(1-b)$ and a variance $v = ab/(1-b)^2$.

As each individual in a sexual contact network as at least one sexual partner, we consider a truncated negative binomial distributions where [23]:

$$Pr^*[k] \;=\; \frac{Pr[k]}{1-Pr[0]} \;=\; \frac{(1-b)^{-a}}{(1-b)^{-a}-1}\binom{a+k-1}{k-1} b^k (1-b)^a \tag{S2}$$

The law has a mean $\mu^* = \frac{ab}{1-b}\,\frac{(1-b)^{-a}}{(1-b)^{-a}-1}$ and a variance $v^* = \frac{ab}{(1-b)^2} + \left(\frac{ab}{(1-b)^2}\right)^2 \frac{(1-b)^{-a}}{(1-b)^{-a}-1} - \left(\frac{ab}{(1-b)^2}\right)^2 \left(\frac{(1-b)^{-a}}{(1-b)^{-a}-1}\right)^2$.

In the context of sexual networks, the $NB(a, b)$ model implies a search mechanism where sexual partners are acquired with probability b until the search is stopped when a suitable partners are found [24]. Here we set the probability that a partner is successfully retained to $b_c = 1 - b = 0.26$ based on surveys of men sexual behaviours in the USA [25].

We generated theoretical negative-binomial networks with the so-called 'configuration model' (CM) [26]. In the CM, we use $Pr^*[k]$ (see equation [2]) with μ^* set to the average number of partners of empirical MSM networks. As b (see above) and $\mu^* = 4$ are known, a is derived immediately.

A unipartite network can be constructed from any size distributions $Pr[k]$ using the configuration model [26]:

1. Each node i is assigned a random number k_i of 'stubs', i.e. ends of edges emerging from the node. k_i is drawn from the degree distribution $Pr[k]$ with $1 \leq k_i \leq N-1$ (a node can not have a degree larger than $N-1$) and imposing the constraint that $\sum_i k_i$ must be even.
2. The network with pre-assigned degree distribution $Pr[k]$ is constructed by connecting pairs of stubs uniformly at random to yield complete edges.

S1.1.7. Randomly generated small world (SW) networks

We generated theoretical SW networks using the Watts-Strogatz model [27] with a rewiring coefficient $r = 0.01$. For the range of nodes explored ($n = 500$ to 4000), $r = 0.01$ exhibit a low average shortest path length as with $r = 1$ (pure random networks) and a high clustering coefficient as with $r = 0$ (pure local networks), thus yielding SW networks exhibiting strong small-world properties [5].

S1.2. Weighting sexual-contact networks

The model we develop allows one to model the spread of a STI (which occurs on a time scale of days) on a sexual contact network (which is built

over a time scale of years). Here, the nodes of the network are individuals and the edges indicate that at least one sexual contact occurred between two individuals over several years (more precisely, over the number of years the network summarises). We consider a static and unipartite network but our results also apply to bipartite and dynamical networks.

S1.2.1. Network topology

The topology describes the way the individuals are connected to each other. All the information on the topology is contained in the adjacency matrix $\mathcal{A}$, whose elements are binaries (0 or 1). The degree of the individuals (i.e. the number of sexual partners) is stored in a degree vector $\mathbf{k}$.

S1.2.2. Individual weights

These weights represent the propensities of individuals to interact with each other. In the case of sexually transmitted infections (STIs), these propensities to interact can be interpreted as potential number of sex acts (PSA) per unit of time. These PSA are contained in a propensity vector $\mathbf{b}$, the ith component of which (b_i) represents the willingness of individual i to interact with other individuals.

PSA are related to 'relational effects' if there is a positive or negative correlation between the weights and the degrees of a node. A positive correlation corresponds to a case where individuals with many PSA tend have the highest number of partners and conversely.

Here, we assume three models of PSA allocation:

1. The linear allocation model (A_{lin}), where $\mathbf{b}$ is strictly proportional to $\mathbf{k}$.

2. The saturating allocation model (A_{sat}, where $\mathbf{b}$ is proportional to $\sqrt{\mathbf{k}}$.
3. The constant allocation model (A_{cst}), where PSA is independent of node degree and $b_i \propto \overline{k}$ ($\overline{k}$ being the average degree of the network).

Classical approaches implicitly assume a linear allocation.

S1.2.3. Edge weights (i.e. realised number of sex acts)

Partitioning potential number of sex acts (PSA). We have to assume a rule for individuals to partition their PSA among their sexual partners. The set of all the partners of individual i is denoted Γ_i.

Availability of a focal node i for a node j. Let $[i \to j]$ be the availability of focal individual i for interacting with individual j. $[i \to j]$ represents the amount of time that individual i can spend with individual j per unit of time given that i is also spending time with his/her other partners, and that these neighbours also spend time with their own partners, etc. If i and j are not connected (i.e. if $j \notin \Gamma_i$), we have $[i \to j] = [j \to i] = 0$. By definition, we also have

$$\sum_{j \in \Gamma_i} [i \to j] = b_i \tag{S3}$$

We consider three cases:

1. Equal partitioning (P_{equ}), where all the partners of an individual i get a share $[i \to j] = b_i/k_i$.
2. Random partitioning (P_{rand}), where all the partners of an individual i get a share $[i \to j] = b_i\,\rho_j/R$, where ρ_j is drawn in a uniform distribution and $R = \sum_{j \in \Gamma_i} \rho_j$.

3. The maximisation partitioning ($P_{\max}$), which is further described below and where individuals use information about the topology to maximise their number of realised sex acts.

Classical approaches assume either a equal or a maximisation partitioning (both are equivalent with a linear allocation model, as we show below).

Edge weights. The 'weights' of the links between individuals represent the frequencies of interactions at each time-step or the number of realized sex acts per time-step for all pairs of individuals in the network.

The weights of links between two individuals i and j (denoted $W(i,j)$) are contained in the weighted adjacency matrix W. If nodes i and j are not connected, $W(i,j) = \mathcal{A}(i,j) = 0$. If nodes i and j are connected, $W(i,j)$ is a function f_{ij} of the propensity vector $\mathbf{b}$ and of the degree vector $\mathbf{k}$. More precisely, $W(i,j) = \mathcal{A}(i,j)\, f_{ij}(\mathbf{b},\mathbf{k})$, which remains true when $\mathcal{A}(i,j) = 0$.

As explained in the main text, we define the interaction function as

$$W(i,j) = \min([i \to j], [j \to i]) \tag{S4}$$

We clearly see that $W(i,j) = W(j,i)$ and that $W(i,j)$ represents the amount of time that individual i and j spend together per time step given that i and j also spend time with their other neighbours, and these neighbours also spend time with their own neighbours, etc. We have the relations:

$$\sum_{j \in \Gamma_i} W(i,j) \leq b_i \tag{S5}$$

$$\sum_{i \in \Gamma_j} W(i,j) \leq b_j \tag{S6}$$

The case of unweighted networks

Modelling STI spread on unweighted networks corresponds to an extreme situation where b_i (i.e. the total number of PSA per unit of time) of an individual i is strictly proportional to its number of sex partners k_i (we have a linear allocation) and where PSA partitioning is equal or maximising. In this case

$$[i \to j] \;=\; \frac{b_i}{k_i} \;=\; \alpha\,\frac{k_i}{k_i} \;=\; \alpha \;=\; [j \to i] \tag{S7}$$

Hence,

$$W(i,j) \;=\; \min(\alpha,\alpha) \;=\; \alpha \tag{S8}$$

It follows from equations S17 and S8 that the total number of sex acts s_i realised by individual i is

$$s_i \;=\; \sum_{j\in\Gamma_j} W(i,j) \;=\; b_i \tag{S9}$$

and that the network is unweighted. We have therefore $s_i \propto k_i$.

S1.2.4. Loss of sex acts due to the weighting

As highlighted above in condition S5, in the general case the availability of i to j is unlikely to be exactly the same as the availability of j to i. This leads to a loss of sex acts, which we quantify on the whole network with a parameter δ, where

$$\delta \;=\; \frac{\sum_i \sum_j W(i,j)}{\sum_i \sum_j \mathcal{A}^*(i,j)} \tag{S10}$$

where the $W(i,j)$ are the terms of the weighted adjacency matrix W and $\mathcal{A}^*(i,j)$ are the terms of the generalised adjacency matrix $\mathcal{A}^*$, which is defined by $\mathcal{A}^*(i,j) = s^*\,\mathcal{A}(i,j)$, where s^* is the average number of realised sex acts

between two individuals (which is set to the same value in all our simulations to get comparable results).

We want to compare the spread of HIV on different weighted networks and this spread increases with the number of sex acts realised on the network. In order to isolate the effect of the network itself, we define the weights on the network such that the average number of sex acts realised per individual per unit of time on a network is equal to the average degree of this network ($\overline{k}$). We hence calculate a corrected weighted adjacency matrix $\tilde{W} = W/\delta$. Note that for the genuine GC network, $\overline{k}_{GC} = 2$ and in order to get values comparable to the other networks (where $\overline{k}_{PL/NB} \approx 4$) we corrected the W for GC by $2\overline{k}_{GC}$.

S1.2.5. Maximisation partitioning

The third partitioning we consider is more complex, which is why we discuss it separately. It assumes that individuals try to maximise their number of realised sex acts by taking into account information about the network topology.

This partitioning can be used with different orders of moment of closure because an individual can partition his/her PSA according to what his/her partners give to him (order 1), but then these partners also have their own partners and can decide to partition their PSA according to what their partners give to them (order 2), etc. At the first order of closure, we close the recursion with the partners j of the focal individual i. In other words, for all the $j \in \Gamma_i$, we assume that

$$[j \to i] \;=\; \frac{b_j}{k_j} \tag{S11}$$

which corresponds to an equal partitioning. Then, the focal will partition his/her PSA b_i among his/her partners j in proportion of what j gave to him/her. Mathematically,

$$\forall j \in \Gamma_i \quad [i \to j] \;=\; b_i \, \frac{[j \to i]}{\sum_{v \in \Gamma_i} [v \to i]} \tag{S12}$$

where Γ_i represents the node-neighbourhood of node i, i.e. the set containing the individuals v connected to i (not including i itself).

At the first order of closure, equation S12 can easily be solved (at least numerically) because

$$\forall j \in \Gamma_i \quad [i \to j] \;=\; b_i \, \frac{b_j/k_j}{\sum_{v \in \Gamma_i} b_v/k_v} \tag{S13}$$

At higher orders of closure, equation S12 is still valid but the availabilities are more complicated. For instance, at order 2, if we still consider a focal individual i, to know $[j \to i]$, we need to know how j, as a focal individual, partitions his/her PSA among his/her partners $l \in \Gamma_j$.

More generally, for a closure moment of order m, the availabilities for any focal individual i can be inferred by recursively with the following equation:

$$\forall m \in \mathbb{N}^\star \quad [i \to j]_{m+1} \;=\; b_i \, \frac{[j \to i]_m}{\sum_{v \in \Gamma_i} [v \to i]_m} \tag{S14}$$

where $[i \to j]_m$ is the availability of i to j obtained at moment closure m, Γ_i represents the node-neighbourhood of node i, i.e. the set containing the individuals v connected to i (not including i itself). Practically, this means that to find all the $[v \to i]_m$, we need to study each neighbour v as a focal and compute all the $[w \to v]_{m-1}$, with $w \in \Gamma_v$. We continue until we reach the $m+1^{\text{th}}$ neighbour of i, denoted x, for which we assume that

$$[x \to y]_1 \;=\; \frac{b_x}{k_x} \tag{S15}$$

where $y \in \Gamma_x$. This assumption of an equal partitioning for the m^{th} neighbour of i allows us to close our recursion.

Note that for individuals with only one partner, i.e. the ends of the network, the partitioning is easy: they attribute all their PSA to their single partner. The problem is that, even with this simplification, there usually is no straightforward way to compute all the availabilities analytically for a closure moment strictly greater than 1 (and even less for an arbitrary moment m). This is why here we use numerical simulations to estimate these availabilities.

The higher the order, the more complete the information is. For the theoretical networks, we could only go up to order 3 because computational time increased exponentially. For the genuine GC network, we were able to go up to order 4.

It is however possible to derive exact analytical solutions from S14 for local networks (see below). Moreover, we notice a convergence of $[i \to j]_m$ when m goes to infinity, which can be written

$$[i \to j] = [j \to i]_\infty = \lim_{m \to \infty} [i \to j]_m \tag{S16}$$

In the term $[i \to j]$, the focal individual is the node i whereas in the term $[j \to i]$, the focal individual is node j. Hence, the neighbours v taken into account in $[i \to j]$ differ from the neighbours v taken into account in $[j \to i]$ and consequently $[i \to j] \neq [j \to i]$ in the general case.

Note that definition S14 respects the normalization condition:

$$\sum_{j \in \Gamma_i} [i \to j] = b_i \tag{S17}$$

If the network is regular, then $k_i = \overline{k}$. With our assumptions, this implies that $b_i = b$. From S14, we thus directly have for this case $[i \to j] = b/k$.

S1.3. Epidemiological modelling

S1.3.1. The SI model

As described in the main text, we assume one of the simplest epidemiological settings, i.e. the SI model [28]. Here, susceptible individuals can become infected when they have sex with infected individuals, with a probability of transmission of β per sex act. Here, we study the spread of a STI on a network, which means that an infected individual can only infect his/her sexual partners. We assume that there is no recovery and that we are studying the population on a short enough time scale so that the population size remains constant (no births nor deaths). We also neglect immigration or emigration in the host population so the networks remains the same.

S1.3.2. Estimating the speed of disease spread

A classical measure of the spread of a disease is the basic reproduction rate (denoted R_0), which estimates the number of secondary infections that an infected host generates if all the other individuals in the population are susceptible [28]. If we take into account the network of host interactions, it is possible to estimate R_0 (even if often sometimes less straightforward).

With non-weighted networks, the speed of disease spread can be estimated through the dominant eigenvalue of the adjacency matrix ($\mathcal{A}$), as shown in [29, 30]. May and Anderson [31, 32] also derived a more general method to estimate the spread of a disease in a heterogeneous population (which is the case of a population on a network). This method is further described below but, in short, it assumes that hosts can be grouped into different transmission groups depending on the number of contacts they have. In a network case,

such an assumption means that we ignore the specificity of the network and only work with the distribution of sexual partners.

When it comes to weighted network, there currently is no specific estimator. This is why we estimate the doubling time from simulations. Using the adjacency matrix to predict disease spread will be the focus of another study.

S1.3.3. Stochastic simulations

We use the Gillespie direct algorithm [33] to run stochastic epidemics on continuous time. For each individual i, the rate corresponding to the event of i becoming infected is the sum of all the βw_{ij}, where j are the sexual partners of i who are infected and β is the disease transmission rate per sex act.

For each network, we infect randomly 1 node and compute the time elapsed to go from 1% of the nodes infected to 2% of the nodes infected (the time elapsed to go from 4 infected nodes to 8 infected nodes for the GC empirical network). This estimator allows us to compare networks of different sizes. Because of the over-dispersion of the data, we use the median doubling time as our estimator and denote it t_d. This definition of the doubling time has the advantage that it removes some of the stochasticity of the very early stages of the epidemics, while remaining an measure of initial growth (which can be associated to R_0).

For any couple allocation-partitioning on the GC network, we generated 100 networks with different weightings and did 1000 runs for each network to estimate t_d. Increasing the number of replicates did not seem to affect the results.

To model prevention policies, we remove one (or multiple nodes) and compute the change on t_d. In the random policy, nodes removed are chosen at random. In the policy targeting individuals with many partners, we remove the most connected nodes first. In cases with random partitioning (P_{rand}), the change is estimated by comparing the t_d for the weighted network with all the nodes and the weighted network without one (or several nodes).

S2. Supplementary Results

S2.1. Effect of prevention policies on theoretical power-law networks

All the results of Figure 5 of the main text are obtained with the genuine GC network. For theoretical power-law (PL) networks with 500 nodes, we find that removing the most connected individuals is the best way to decrease the disease doubling time (Figure S2). However, the efficiency of this strategy varies with the assumptions made to weight the network. A generalised linear model yielded significant correlations with a slope of 0.19 for the case with linear allocation with equal partitionning, 0.16 for the same allocation but with random partitionning and 0.08 for the case with constant allocation of potential number of sex acts (PSA) and random partitionning. Also, in the latter case, the nodes that have a largest effect on disease spread have a intermediate degree.

S2.2. Removing multiple nodes

Here, we study a case where many individuals can be removed. This removal can be done in three ways: i) randomly, ii) following node degree (i.e. target first highly connected individuals) and iii) following sex acts (i.e. target first individuals with the higher number of realised sex acts).

Figure S3A shows that increasing the number of nodes removed tends to increase the median doubling time of a disease spreading on the GC network weighted assuming linear PSA allocation and random partitionning. However, there is a decrease in median doubling time if 7 individuals are removed. This is likely to be due to specificities of the GC network because we do not see such a decrease for theoretical PLnetworks (see below). Also, if many

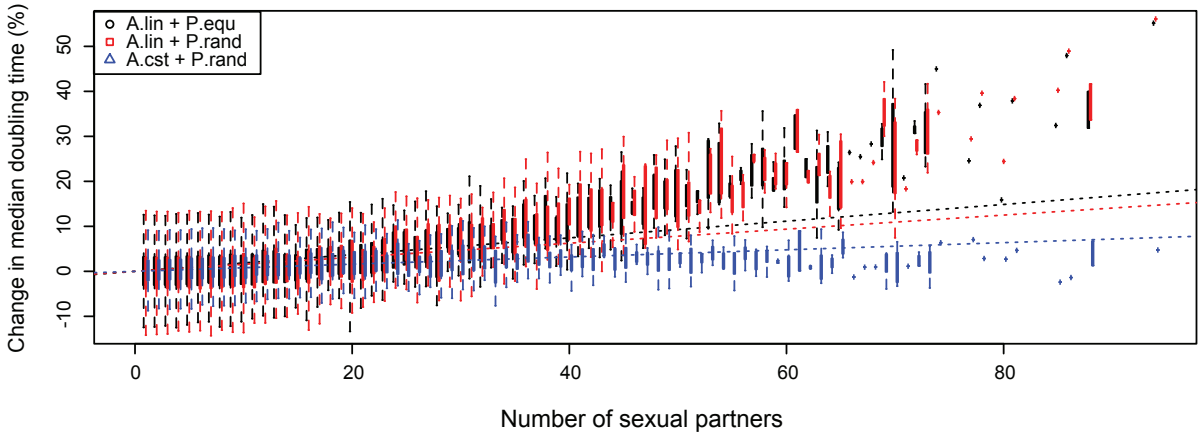

Figure S2: Change in median doubling time after node removal as a function of node degree for different PSA allocations

In black we have linear allocation and equal partitionning, in red linear allocation and random partitionning and in blue constant allocation and random partitionning. Box plots shows median values, the three quartiles and the outliers for 100 theoretical PL network with 500 nodes with 10000 runs per network. Here, $\beta = 0.01$ and individuals have on average 4 sex acts per week.

nodes can be removed, targeting the most highly connected individuals (in black) tends to be the best strategy. This is shown on panel B: as soon as at least 3 individuals are removed, targeting the most connected individuals (or the individuals with the highest number of sex acts) leads to a significantly more important decrease in doubling time than random targeting (see the figure caption for further details about the statistical test used).

The same approach can be applied to the GC network weighted assuming constant PSA allocation and random partitionning. In this case, the consequence of node removal on doubling time is much weaker (figure S3C).

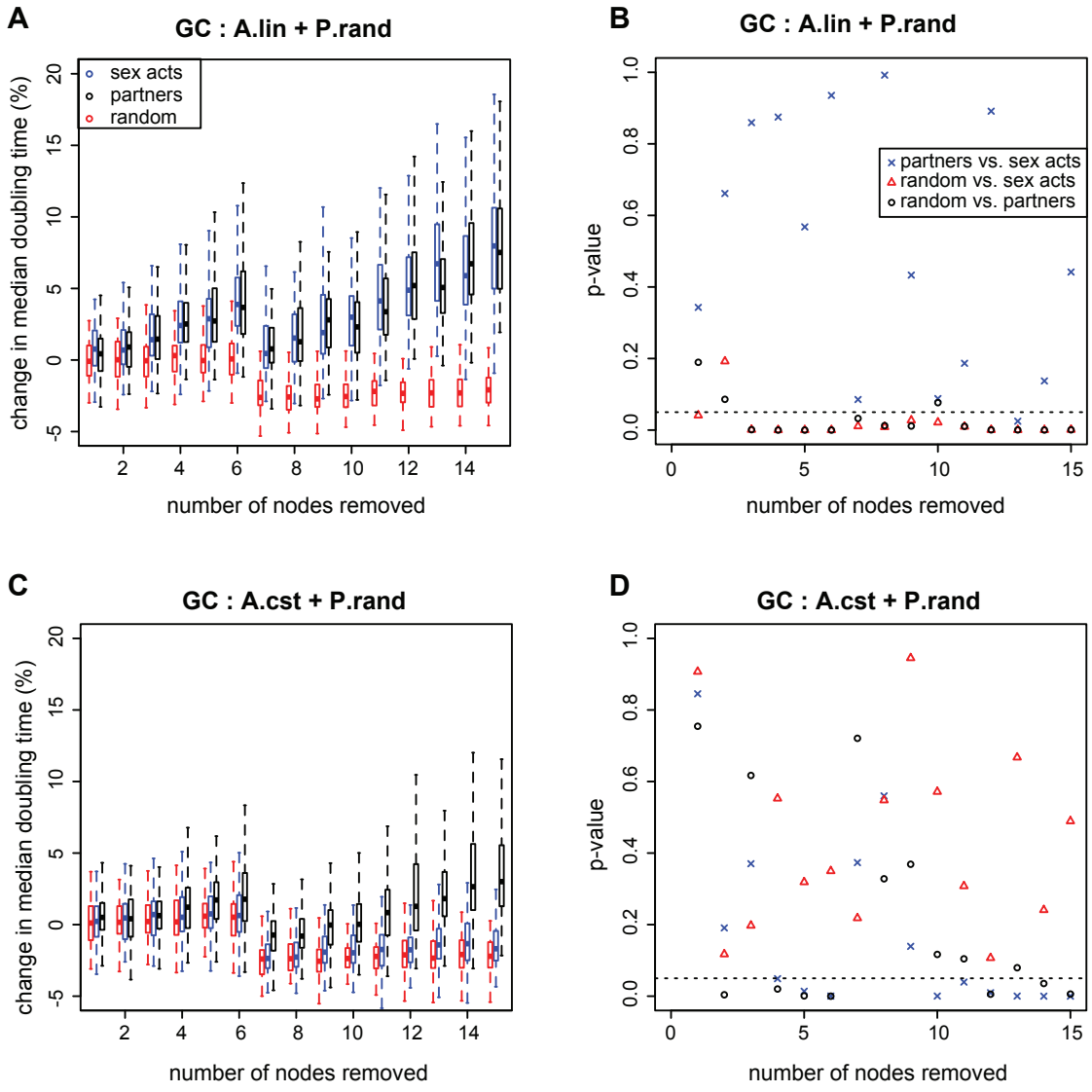

Figure S3: Consequence of multiple node removal on median doubling time for the GC network

Panels A and C show the change in median doubling time as a function of the number of nodes removed. The colour indicates the strategy of node removal (red is random, blue is targeting for sex acts and black is targeting for partners). Panels B and D show the p-value of a Wilcoxon signed-rank test, which tests for an effect of the removal strategy on the median doubling time. The colors indicate the strategies that are compared (see the caption on panel B). In A and B the network is weighted using a linear allocation and a random partitionning. In C and D the network is weighted using a constant allocation and a random partitionning. Here, $\beta = 0.01$ and individuals have on average 4 sex acts per week. Box plots shows median values, the three quartiles and the outliers for the median t_d.

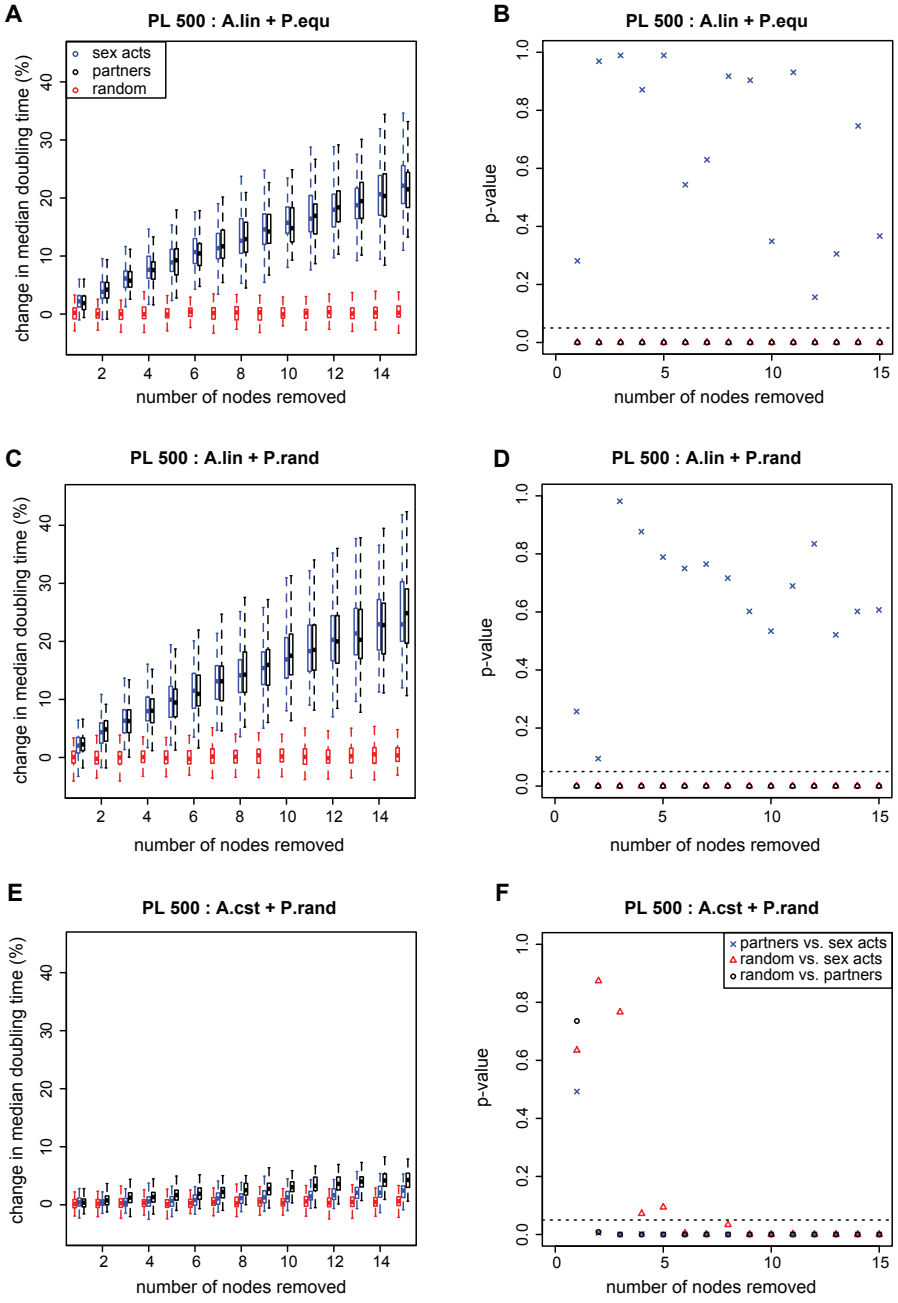

Figure S4: Consequence of multiple node removal on median doubling time for theoretical PL network

In A and B the network is weighted using a linear allocation and an equal partitionning. In C and D the network is weighted using a linear allocation and a random partitionning. In E and F the network is weighted using a constant allocation and a random partitionning. Here, $\beta = 0.01$ and individuals have on average 4 sex acts per week. Box plots shows median values, the three quartiles and the outliers for 100 theoretical PL networks with 500 nodes with 1000 runs per network.

Moreover, it takes many nodes to be removed before one can see a different between the strategies (figure S3D).

In order to check that these results were not only due to the specificity of the GC network, we applied a similar approach to 100 theoretical PL networks of 500 nodes (figure S4). In this case we had enough data to also consider the case of a linear allocation with an equal partitionning (figure S4A and B). In this case, as for the case with linear allocation and random partitionning (figure S4C and D), we see a significant difference between the targeting strategies as soon as one node is removed. In the case of the constant allocation however, it takes 2 individuals to be removed before a significant difference can be observed between the random targeting and the targeting for individuals with high number of partners (figure S4E and F). It even takes 4 individuals to be removed if the targeting is based on sex acts. Finally, the magnitude of the change in doubling time is much smaller in the constant allocation case, suggesting that targeted policies are less efficient.

These results illustrate that taking into account the number of sex acts explicitly leads to network topologies where prevention strategies aimed at the most connected individuals require to target a larger subset of the population to be more efficient than policies with random targeting.

S2.3. Doubling times on the GC network

We here present in details the results of the generalised linear model tests assessing the effect of key parameters on median doubling times on the GC network (Table 3).

Table 3: Effect of biological assumptions on the observed median doubling time in the GC network

We use a generalised linear model to test for differences with the case with a linear allocation and an equal partitionning. Significant effects are in bold font.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.7895	0.0524	53.21	<0.001
replicate	0.0001	0.0001	0.94	0.348
allocation sat	0.0788	0.0741	1.06	0.289
allocation **cst**	**0.1471**	0.0741	1.98	**0.048**
partitionning max	0.0000	0.0741	0.00	1.000
partitionning **rand**	**0.1042**	0.0529	1.97	**0.050**
sat : max	0.1965	0.1049	1.87	0.062
cst : max	**0.5428**	0.1049	5.18	**<0.001**
sat : rand	-0.0021	0.0745	-0.03	0.977
cst : rand	0.0238	0.0745	0.32	0.749

S2.4. Results for networks with a power-law (PL) node degree distribution

All the results of Table 1 in the main text were also checked on theoretical power-law networks. Here, we present the results of simulations done on all the networks (genuine and theoretical are pooled into a single data set). For each of the networks (GC, PL with 500 nodes and PL with 4000 nodes) we studied two transmission rates ($\beta = 0.01$ and $\beta = 0.001$) We study the effect of allocation model, partitionning model, network size and transmission rate (β) on observed median doubling times.

When all the networks are pooled we find, as for the GC network alone, that constant allocation and random partitionning lead to doubling times that differ from the classical approach (Table 4). We also find here that the saturating allocation has a significant effect. Since we have different types of network, we can study the effect of the size (N) of the network and of the transmission rate (β): both affect the doubling time. When we look at interaction terms, we find that there is a strong interaction between the network size and the allocation model or the random partitionning model. Note that all the effects on the doubling time of the network size were weak (lower than 1e-5 for weighted networks). This could be linked with the effect that we describe in the main text: doubling time is expected to decrease with N on PL networks and the weighting can cancel this effect. This interpretation is supported by the fact that we find no significant interaction involving the transmission rate, which is only a scaling parameter.

Table 4: Effect of biological assumptions on the observed median doubling time on all the networks

None of the interaction between more than two factors were significant.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.2108	0.0247	89.65	<0.001
replicate	-0.0001	0.0001	-1.24	0.214
allocation **sat**	**0.1809**	0.0347	5.22	**<0.001**
allocation **cst**	**0.3354**	0.0347	9.68	**<0.001**
partitionning max	-0.0000	0.0347	-0.00	1.00
partitionning **rand**	**0.4732**	0.0299	15.81	**<0.001**
N	**3.7e-5**	0.0000	-4.23	**<0.001**
β	**-112.67**	3.4485	-32.67	**<0.001**
sat : max	**0.1315**	0.0490	2.68	**0.0073**
cst : max	**0.2274**	0.0490	4.64	**<0.001**
sat : rand	-0.0664	0.0423	-1.57	0.1167
cst : rand	**-0.1027**	0.0423	-2.43	**0.015**
sat : N	**3.7e-05**	1.2e-5	3.01	**0.0026**
cst : N	**4.8e-5**	1.2e-5	3.90	**<0.001**
max : N	0.0000	0.0000	0.00	1.00
rand : N	**-0.0001**	1.1e-5	-9.87	**<0.001**
sat : β	1.6143	4.8769	0.33	0.741
cst : β	1.2468	4.8769	0.26	0.798
max : β	0.0000	4.8769	0.00	1.00
rand : β	1.5930	4.2129	0.38	0.705
N : β	0.0004	0.0012	0.33	0.744

S2.5. Results for small-world (SW) networks

Table 5: Effect of network size on disease spread for three allocation models. For further details, see the caption of figure S5.

	$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$
P_{rand} A_{lin}	68 (46-99)	129 (92-177)	211 (149-295)	274 (206-387)
P_{rand} A_{sat}	69 (47-100)	129 (94-177)	210 (150-294)	277 (207-386)
P_{rand} A_{cst}	68 (47-99)	131 (95-177)	214 (149-296)	281 (213-389)

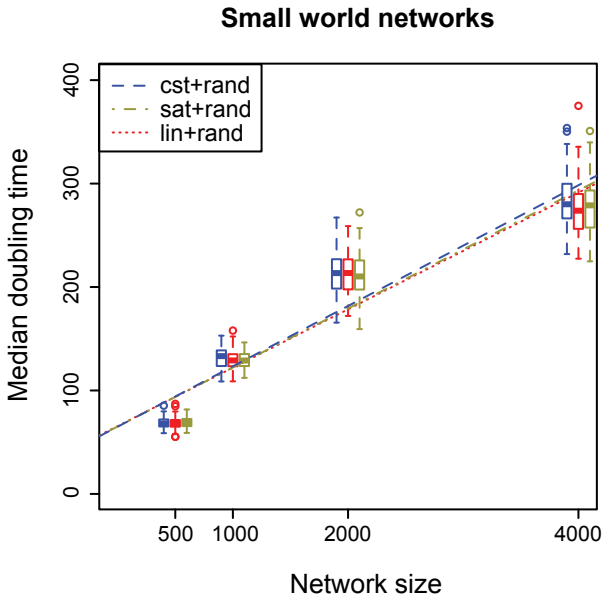

Figure S5: Doubling time for theoretical SW networks of different sizes. We weighted theoretical small-world networks using a random partitioning (P_{rand}) with a linear (A_{lin}, in red), saturating (A_{sat}, in brown) or constant (A_{cst}, in blue) allocation. Disease spread is estimated with the median doubling time (t_d) computed over 100 runs of the simulation per replicate network. The box plot shows the median values over 100 different networks, the three quartiles and the outliers. Dashed lines are the output of a generalised linear model testing the effect of network size on t_d (see the main text). Parameters are as in Figure 2 of the main text

[1] S. H. Strogatz, Exploring complex networks, Nature 410 (6825) (2001) 268–276, doi:10.1038/35065725.

[2] M. E. J. Newman, The structure and function of complex networks, SIAM Review 45 (2) (2003) 167–256.

[3] M. E. J. Newman, Spread of epidemic disease on networks, Phys. Rev.

E 66 (1) (2002) 16128, doi:10.1103/PhysRevE.66.016128.

[4] M. J. Keeling, K. T. Eames, Networks and epidemic models, J. R. Soc. Interface 2 (4) (2005) 295–307, doi:10.1098/rsif.2005.0051.

[5] D. Watts, Small worlds: the dynamics of networks between order and randomness, Princeton University Press, 1999.

[6] J. Gómez-Gardeñes, V. Latora, Y. Moreno, E. Profumo, Spreading of sexually transmitted diseases in heterosexual populations, Proc. Natl. Acad. Sci. USA 105 (5) (2008) 1399–404, doi:10.1073/pnas.0707332105.

[7] C. Bauch, D. A. Rand, A moment closure model for sexually transmitted disease transmission through a concurrent partnership network, Proc. R. Soc. Lond. B 267 (1456) (2000) 2019–2027, doi:10.1098/rspb.2000.1244.

[8] L. A. Meyers, M. E. J. Newman, B. Pourbohloul, Predicting epidemics on directed contact networks, J. theor. Biol. 240 (3) (2006) 400–418, doi:10.1016/j.jtbi.2005.10.004.

[9] M. J. Jeger, M. Pautasso, O. Holdenrieder, M. W. Shaw, Modelling disease spread and control in networks: implications for plant sciences, New Phytol 174 (2) (2007) 279–97, doi:10.1111/j.1469-8137.2007.02028.x.

[10] A. L. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512, doi:10.1126/science.286.5439.509.

[11] M. L. Goldstein, S. A. Morris, G. G. Yen, Problems with fitting to the power-law distribution, European Physical Journal B 41 (2) (2004) 255–258, doi:10.1140/epjb/e2004-00316-5.

[12] B. Bollobás, C. Borgs, J. Chayes, O. Riordan, Directed scale-free graphs, in: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, SIAM, PA, USA, 132–139, 2003.

[13] C. Cooper, A. Frieze, A general model of web graphs, Random Struct. Algor. 22 (3) (2003) 311–335, doi:10.1002/rsa.10084.

[14] J. J. Potterat, R. B. Rothenberg, D. E. Woodhouse, J. B. Muth, C. I. Pratts, J. S. Fogle, Gonorrhea as a Social Disease, Sex. Transm. Dis. 12 (1) (1985) 25–32.

[15] J. J. Potterat, S. Muth, H. Stites, Twenty-year mortality in a 1981 cohort of homosexuals with gonorrhoea: a preliminary estimate, Int J STD AIDS 12 (6) (2001) 413–416.

[16] B. Bolker, Maximum likelihood estimation and analysis with the bbmle package, R package. Version 0.8 .

[17] F. Liljeros, Sexual networks in contemporary Western societies, Physica A 338 (1-2) (2004) 238 – 245, ISSN 0378-4371, doi: 10.1016/j.physa.2004.02.046.

[18] M. E. J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA 103 (23) (2006) 8577–8582, doi: 10.1073/pnas.0601602103.

[19] A. M. Johnson, C. H. Mercer, B. Erens, A. J. Copas, S. McManus, K. Wellings, K. A. Fenton, C. Korovessis, W. Macdowall, K. Nanchahal, S. Purdon, J. Field, Sexual behaviour in Britain: partnerships,

practices, and HIV risk behaviours, Lancet 358 (9296) (2001) 1835–42, doi:10.1016/S0140-6736(01)06883-0.

[20] N. Bajos, M. Bozon, N. Beltzer, Enquête sur la sexualité en France: pratiques, genre et santé, La Découverte, Paris, 2008.

[21] I. J. Farkas, I. Derenyi, A. L. Barabasi, T. Vicsek, Spectra of 'real-world' graphs: Beyond the semicircle law, Phys. Rev. E 6402 (2) (2001) 026704, doi:10.1103/PhysRevE.64.026704.

[22] R. M. May, Network structure and the biology of populations, Trends Ecol. Evol. 21 (7) (2006) 394–399, doi:10.1016/j.tree.2006.03.013.

[23] N. Johnson, S. Kotz, A. Kemp, Univariate discrete distributions (2nd edition), 1992.

[24] D. T. Hamilton, M. S. Handcock, M. Morris, Degree distributions in sexual networks: a framework for evaluating evidence, Sex. Transm. Dis. 35 (1) (2008) 30–40, doi:10.1097/OLQ.0b013e3181453a84.

[25] M. S. Handcock, J. H. Jones, Interval estimates for epidemic thresholds in two-sex network models, Theor. Popul. Biol. 70 (2) (2006) 125–34, doi:10.1016/j.tpb.2006.02.004.

[26] M. Catanzaro, M. Boguñá, R. Pastor-Satorras, Generation of uncorrelated random scale-free networks, Phys. Rev. E 71 (2) (2005) 027103, doi:10.1103/PhysRevE.71.027103.

[27] D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440–2, doi:10.1038/30918.

[28] R. M. Anderson, R. M. May, Infectious diseases of humans: dynamics and control, Oxford University Press, New-York, USA, 1991.

[29] J. G. Restrepo, E. Ott, B. R. Hunt, Characterizing the dynamical importance of network nodes and links, Phys. Rev. Lett. 97 (9) (2006) 094102, doi:10.1103/PhysRevLett.97.094102.

[30] D. Chakrabarti, Y. Wang, C. X. Wang, J. Leskovec, C. Faloutsos, Epidemic thresholds in real networks, ACM Transactions on Information and System Security 10 (4) (2008) 1, doi:10.1145/1284680.1284681.

[31] R. M. May, R. M. Anderson, Transmission dynamics of HIV infection, Nature 326 (6109) (1987) 137–42, doi:10.1038/326137a0.

[32] R. M. May, R. M. Anderson, The transmission dynamics of human immunodeficiency virus (HIV), Philos. Trans. R. Soc. B 321 (1207) (1988) 565–607, doi:10.1098/rstb.1988.0108.

[33] M. J. Keeling, P. Rohani, Modeling infectious diseases in humans and animals, Princeton University Press, Princeton, USA, 2008.