Information processing in complex networks

Quax, R.

Publication date
2013

Chapter 4

Inferring epidemiological parameters from phylogenetic information for the HIV-1 epidemic among MSM

Material from this chapter is published in the European Physics Journal in the Special Topics section. In addition, a technical description of the optimized computing tool SEECN was published in the International Journal for Multiscale Computational Engineering.

4.1 Introduction

Each person with an HIV infection carries a pool of viral genotypes which evolves under pressure from the immune system and treatment. The receiver of a transmitted infection starts with a copy of the viral gene pool of the sender, but gradually develops his own unique genotype through genetic drift. As a consequence, the topology, timing and ordering of infection transmissions in a population leave their fingerprint in the phylogeny of the virus in the population. The goal of our work is to measure how much information is contained in the phylogenetic data about the epidemiological process that created it.

It is difficult to measure epidemiological parameters directly from phylogenetic data. The first problem is that the sequence data is necessarily incomplete. Roughly one quarter of the infected men-who-have-sex-with-men (MSM) is undiagnosed (199, 200), and roughly one third of the diagnosed MSM has not (yet) been sequenced (201). The second problem is that the sequence data is inherently ambiguous about the underlying route of

transmission. In principle, if A and B are two similar sequences, then the transmission event could be either $A \to B$, $B \to A$, or $C \to A$ and $C \to B$. Another source of ambiguity is that HIV-patients are sequenced only once or a small number of times in their life, whereas the period of infectiousness is very long and the rate of genetic mutation is relatively high. This means that even if A transmitted the infection to B, their sequences may not be similar because they were diagnosed and sequenced at different times.

Most previous studies use a panmictic model with one or a few parameters to characterize the underlying transmission dynamics (202). These parameters are then directly estimated from the sequence data. In the classical coalescent methods, a typical set of parameters is a reproduction number and a rate of genetic drift. Such methods are primarily used for their mathematical convenience rather than their faithful description (203, 204). The current trend is to increase the complexity of the epidemiological model step by step, such as adding a death rate (204) or a variable population size (202). Leigh Brown et al. (201) propose to infer the network structure of the HIV transmission by placing a connection between every pair of sequences that has a genetic distance below a cut-off value.

In contrast to this trend, we propose to start at the other end of the complexity spectrum: simulating a detailed model to estimate the likelihood that a given set of parameters would reproduce the observed phylogenetic data. The better a set of parameter values is capable of reproducing the observed phylogenetic data, the more likely the set of parameters describes the underlying epidemiological process. Our research question is whether phylogenetic data is capable of providing significant information about the set of epidemiological parameters in this manner.

Our starting point is the present knowledge of epidemiological parameters, which include the topology and frequency of sexual interactions, the per-act infection probability for all stages, and the risk behavior reduction upon diagnosis. For each parameter we take the best estimate from the literature as well as the uncertainty about the value, in the form of a confidence

interval or a standard deviation. This knowledge typically comes from cohort studies and health reports.

The complexity of our simulations is `data-driven'. In other words, the list of available parameter estimates induces the possible internal states of MSM and the network topology that we model. Our simulations consist of 6000 `agents' connected by a dynamic complex network, where each agent has an individual internal state and behavior. In time steps that are equivalent to 3 months, agents create new sexual contacts with other agents and remove old contacts. Each serodiscordant sexual contact transmits the virus with some probability, and an infected agent individually progresses through the stages acute, asymptomatic, diagnosed, treated, and AIDS. Additionally, agents on treatment may develop a drug-resistant mutation which could be transmitted to others.

We simulate this model many times using the Monte Carlo method where we start each simulation with semi-random parameter values, induced by the present knowledge from the literature. Each sampled set of parameters is then scored with the likelihood that it would reproduce the observed cluster size distribution by Brown et al. (201) from 14560 subtype-B sequences from 2001 through 2007 from the UK HIV Drug Resistance Database (205). We calculate this likelihood by first selecting all agents that were newly diagnosed during the equivalent time span of 2001 through 2007. Then we let two agents cluster together only if they receive the virus from a recent common ancestor, or if they infected each other.

The result is an estimated probability distribution of the epidemiological parameters based on the observed cluster-size distribution. This probability distribution encodes knowledge about the parameters, which may be combined with the existing knowledge from the literature. To quantify how much information is contained in the cluster-size distribution we use Shannon's information theory (84).

4.2 Materials and methods

4.2.1 Current knowledge about the epidemiological parameters

The current best knowledge about the epidemiological parameters is encoded as a collection of best estimates of all parameter values together with their uncertainty, which is scattered across the literature. We gathered these data as best we could from a variety of sources, including cohort studies, health reports, and questionnaires. For each parameter value we defined an appropriate probability distribution based on its expected value and the reported confidence interval or standard deviation. The parameter values are summarized in Figure 10 and Table 1, Table 2, and Table 3.

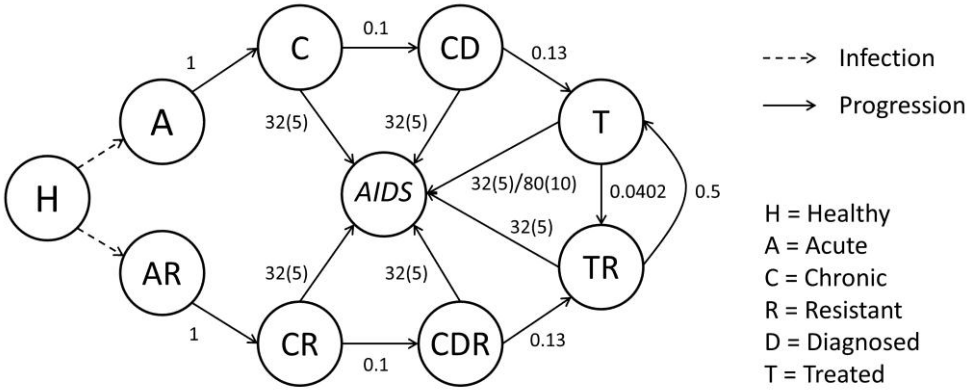

Figure 10: All possible internal states of each MSM and their transitions. A single number r denotes a geometric rate and a pair of numbers $\mu(\sigma)$ denotes a normal distribution.

The uncertainty of each parameter value is encoded as a normal distribution when an appropriate standard deviation could be calculated, and as a geometric rate if not. A geometric rate p implies a skewed probability distribution with standard deviation $\sqrt{1-p}/p$ across individuals. The exception to using geometric rates are the progression times to AIDS in Table 2; here a standard deviation is unknown, but a geometric rate would

imply an inappropriate probability distribution. This is because a progression time to AIDS of one or two years is never observed, but a geometric distribution would imply maximum probabilities for such short progressions. Therefore we assume that the onset of AIDS is a Poisson process, which implies that the standard deviation is the root of the mean.

All parameter values are generated independently at the start of each simulation. This means, for example, that although we impose *on average* a zero effect of drug resistance on the infection probability, in a particular simulation there may be a positive or negative correction factor. As a consequence we also account for the uncertainty of how parameter values relate to each other. In this example we account for the possibility that a drug resistant virus strain is less fit for transmission, for which no conclusive data exists.

For some parameter values we do not model uncertainty if it is already accounted for in another parameter value. For instance, the per-act infection probabilities have very high uncertainty, namely a standard deviation of roughly half the mean, due to the difficulty in their estimation. Therefore we set the frequency of condom use to a constant 50% because its estimated value and range varies significantly in the literature, and let is uncertainty be captured by the infection probability parameter. Another example is the change of behavior upon diagnosis. We use an average 25% reduction of infectiousness, but as a result of the independent uncertainty of both infection probabilities the standard deviation of this parameter is approximately 28 percentage points.

Progression	Rate (per 3 months)	Description
$A \to C$	1	
$C \to D$	$1/(0.47 \cdot 4)$	Median time since seroconversion at diagnosis is 170 days.
$D \to T$	$1/(1.882 \cdot 4)$	Median time between diagnosis and treatment was 687 days in the Erasmus Medical Center in Rotterdam up to 2011.
$T \to TR$	0.0402313	
$TR \to T$	$1/(0.5 \cdot 4)$	We assume semi-annual check-ups.

Table 1: Rates of progression between internal states of persons other than the AIDS stage. The state symbols are defined in Figure 10. The source references are (206–208).

4.2.2 Simulating the HIV epidemic among MSM

We model the epidemic in a data-driven manner. In other words, the complexity of our model is defined by the set of estimated parameter values that is available. This provides our best estimate of the effect of the parameters on the transmission process of HIV among MSM, and consequently the cluster-size distribution that they induce. For example, there is sufficient data about the individual progression rates and infectiousness of HIV-infected persons from the acute phase, which lasts about three months, to the extended asymptomatic period, and finally to the AIDS stage. Consequently we distinguish between these disease stages and model individual agents explicitly. As another example, there is insufficient data available to model the genetic dynamics of the HIV-virus in more detail within individuals, which include genetic drift under pressure of the immune system, the additional effects of the treatments, and the fitness for transmission of different genotypes. As a result we do not distinguish between genotypes of the HIV-virus in our model.

For our experiments we use the SEECN simulation program (209) to simulate the individual behavior of agents which are connected by a dynamical complex network. The agents correspond to MSM and the connections in the network correspond to sexual interactions. In discrete time steps, each agent adds and removes connections according to his inherent promiscuity, which is specified as the expected number of connections of the agent at any time. The internal state of each agent may change due to the influence of connected agents, which corresponds to infection transmission, or due to internal progression, such as becoming treated or progressing to the AIDS stage. See Figure 11.

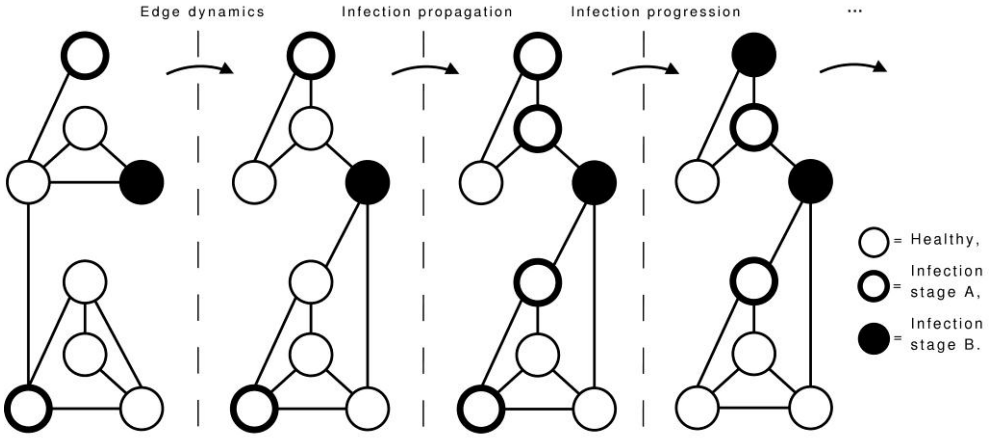

Figure 11: We model individual agents which can infect other agents, progress in stages of disease, and add or remove their connections.

The global topology of the sexual contacts in the MSM population is found to be highly skewed (210), with many individuals having a low promiscuity and only a few individuals having a high promiscuity. We model this skew by imposing a power-law distribution of the expected number of connections of individuals, with an exponent uniformly distributed between 1.5 and 2.0 (210). Although the number of connections, or degree, of an agent may vary over time, the long-term average degree of a MSM is determined by his individual inherent promiscuity. The network topology is

random in any other respect, i.e., we do not impose additional constraints such as assortativity or community clustering.

The internal states of agents as well as the possible paths of progression are shown in Figure 10. At the start of each simulation we set 1% of the agents in the asymptomatic infection stage and the rest in the healthy state, which we assume corresponds to the year 1983 (211). Infected agents that are unaware of their infection will have sexual contacts with other agents as if they were healthy, possibly transmitting the virus. Simultaneously, these unaware infected agents become diagnosed with their condition at a given rate, after which they become less infectious due to changing their behavior (208). Diagnosed agents may become treated, which further reduces their infectiousness due to suppression of the virus. We divide this reduction factor into a pre-HAART and a post-HAART episode because of the significant improvement of HAART (212–214). Sometimes a treated agent develops a drug-resistant mutation, which increases his infectiousness until a different treatment is prescribed. This mutation may be onward transmitted. Lastly, infected agents may progress to AIDS at a different rate for the pre-HAART and post-HAART episode.

Progression	μ	σ	**Description**
$C, CD, CR \rightarrow AIDS$	40	$\sqrt{40}$	Incubation period of AIDS is modeled as a Poisson process with a mean of 10 years.
$T, TR \rightarrow AIDS$	80	$\sqrt{80}$	Pre-HAART. Poisson process.
$T, TR \rightarrow AIDS$	144	$\sqrt{144}$	Post-HAART. Poisson process.

Table 2: Durations of progression to the AIDS state from the various internal states. For each person the duration is generated by a normal distribution with the mean and standard deviation specified in this table. The state symbols are defined in Figure 10. The source references are (208, 215, 216).

4.2.3 Phylogenetic data

Hospitals determine the viral genotypes of all newly diagnosed HIV-infected patients and store them in a database. Each genotype is a sequence of nucleotides, which are symbolized as A, C, G, and T. For each pair of patients we can calculate the similarity of their sequences by counting the number of genes that match. Due to genetic drift of the virus population under pressure of the immune system within each patient, this similarity of genotypes is a measure of how long ago their infection descended from a common ancestor. A high similarity indicates either a short infection route from one patient to the other, or a recent sexual partner who infected them both. A low similarity could indicate an indirect relationship through a long chain of infections, or a time of infection transmission that is long ago. As a consequence the distribution of the similarities of the genomic sequences are somehow a representation of how the HIV-virus was transmitted through the network of sexual contacts in a population.

Unfortunately it is not possible to observe the history of transmissions of HIV directly from this phylogenetic data. The first reason is that it is inherently ambiguous about the underlying causality. Even a very high similarity between the genotypes two patients does not distinguish between a direct infection transmission from one patient to another or vice versa, and a common sexual partner who recently infected both. A low similarity adds more ambiguity because additional routes of infection transmission become possible. It also cannot rule out a direct relationship between the two patients because the time of infection transmission could be a long time ago.

The second reason is that phylogenetic data of the HIV virus is necessarily incomplete. Approximately one quarter of the infected MSM is undiagnosed (199, 200), and one third of the diagnosed MSM has not (yet) been sequenced (201), which means that their viral genotypes are missing. Additionally, sequences of a single person are sparingly sampled which further increases the ambiguity in phylogenetic analyses. As a consequence, the phylogenetic data cannot distinguish between direct or indirect routes of

infections, and is uncertain about the directionality.

HIV stage	$\mu \times 10^{-3}$	$\sigma \times 10^{-3}$	Description
Acute	33.97	14.61	During the first 3 months of infection the infectiousness is an expected 7.25 times higher than in the asymptomatic stage (214).
Acute, resistant	33.97	14.61	No expected effect from resistance.
Asymptomatic	4.685	2.015	
Asymptomatic, resistant	4.685	2.015	No expected effect from resistance.
Asymptomatic, resistant, diagnosed	3.514	1.511	The expected effect of diagnosis is a 25% reduction of infectiousness due to a change in behavior.
Asymptomatic, diagnosed	3.514	1.511	The expected effect of diagnosis is a 25% reduction of infectiousness due to a change in behavior.
Asymptomatic, treated pre-HAART	2.636	1.133	Pre-HAART treatment reduces infectiousness by an expected 25%.
Asymptomatic, treated post-HAART	0.1406	0.06044	HAART reduces infectiousness by an expected 96%.
AIDS			MSM in the AIDS stage do not have sexual contacts in our model.

Table 3: The probability that an MSM transmits the infection depending on his HIV-stage, rounded to four significant digits. All infection probabilities include a correction factor for condom use, for which we use a constant frequency of condom use of 50% and a reduction of per-act infectiousness of 87% (4.9%) (217). The source references for the per-act infection probabilities are (208, 212–214, 218).

In our analysis we use the cluster-size distribution by Brown et al. (201) of the sequence data from the UK HIV Drug Resistance Database (205), which is summarized in Figure 12. This means that we only consider the high similarities that indicate a direct sexual contact between patients, and discard pairs of sequences with lower similarities. The reason is that there is insufficient data available to model the genotype of HIV-virus within individuals, including genetic drift under pressure of the immune system, the additional effects of the treatments, and the fitness for transmission of different genotypes. We can, however, model the direct sexual contacts between individuals that shape the cluster size distribution. The dataset consists of 14560 subtype B sequences of HIV-infected persons in the years 2001 through 2007, of which roughly 80% belong to MSM.

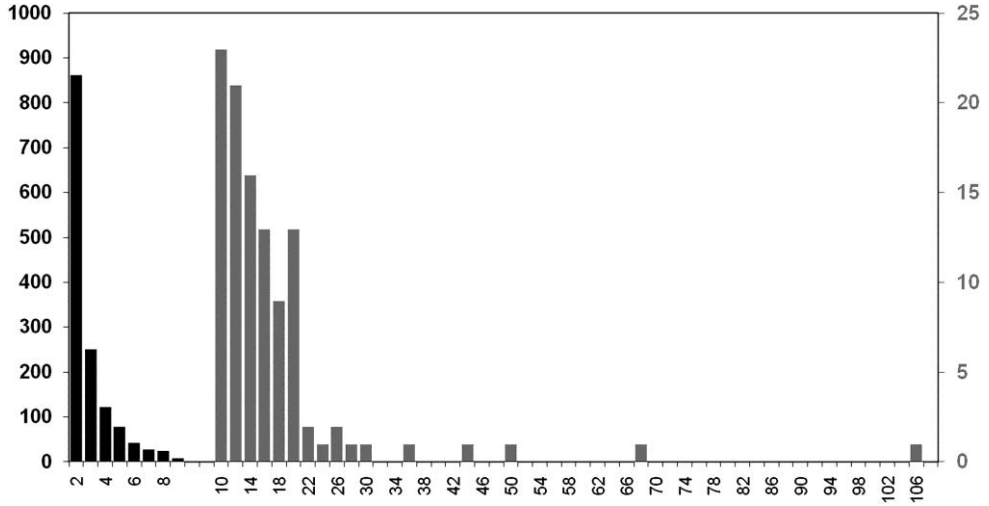

Figure 12: Cluster-size distribution of sequences from the UK HIV Drug Resistance Database (205) calculated by Brown et al. (201). Reprinted with permission.

4.2.4 Calculating the likelihood of reproducing the cluster-size distribution

Each simulation produces a list of infection events between persons in each time step. These infection events form clusters. We select the newly

diagnosed agents in the time steps that correspond to the years 2001 through 2007. Two of these agents cluster together either if they infected each other or the agents that infected them belong to the same cluster, i.e., share a recent common ancestor in the chain of infection events.

Of the resulting cluster-size distribution we calculate the likelihood that it could be a subset of the observed cluster-size distribution. The reason for this is that the size of our simulations is much smaller than the number of MSM in the United Kingdom. Suppose that the number of MSM in the simulated cluster-size distribution is N_C. If we randomly select N_C individuals from the UK HIV Drug Resistance Database sequences then there are many possible subsets, each with a corresponding cluster-size distribution. Each such subset cluster-size distribution may be obtained via multiple selections. The likelihood of a given subset cluster-size distribution is proportional to the number of selections of sequences that lead to it.

More precisely, the expected number of clusters c_s of size s in the subset is

$$E[c_s] = \sum_{s'=s}^{106} \text{Binom}[s', N_C / 14560] \cdot \left|C_{s'}\right|, \tag{3}$$

where $\text{Binom}[n, p; k]$ is the binomial probability distribution, and $\left|C_{s'}\right|$ is the number of clusters of size s' that were observed. We approximate the variability of the number of s-clusters with a Poisson process, i.e.,

$$\sigma^2[c_s] = E[c_s]. \tag{4}$$

This yields our estimate of the probability that the subset cluster-size distribution has c_s clusters of size s. The probabilities for all possible cluster sizes are multiplied, which is equal to the likelihood.

In reality, two individuals would not cluster together if their most recent common ancestor is too long ago, due to genetic drift. Therefore we only let two agents from our simulations cluster together if their most recent

common ancestor (indirectly) infected both agents after the year 1998 in simulation time, which is 3 years before the start of the sequences dataset. This conservatively excludes most clustering sequence pairs which are biologically unlikely, but it does not exclude all. This is not possible because little data exists on the time it takes for two individuals to change their genetic sequence such that they would no longer cluster in a phylogenetic analysis. Fortunately, we are not interested in the absolute likelihood of a particular simulated cluster-size distribution; we require only the ratio of two likelihoods. If each likelihood has a similar multiplicative bias due to false clustering pairs, then the normalized distribution of the likelihoods is not affected by the bias.

4.3 Results

Our main finding is that the cluster-size distribution indeed contains substantial information about the underlying epidemiological parameters. Here we show this for three important epidemiological parameters: the network topology of the sexual interactions, the per-act infection probability, and the change of behavior upon the diagnosis of HIV. See Figure 13.

As before, we quantify the amount of information using Shannon's formula of entropy in the form of mutual information (84). For a parameter X with possible values $1,...,n$, the number of bits required to determine the value of X (without further knowledge) equals

$$H[X] = \log_2 n .$$

A probability distribution p_i over these possible values contains information about the value of the parameter. Intuitively, a shallow distribution contains little information because many values would still be possible. A sharp distribution, on the other hand, contains a lot of information because it rules out most values and narrows down to only a few possibilities. Quantitatively, a probability distribution p_i yields

$$I[X] = H[X] - H[X \mid p_i]$$

bits of information because it reduces the uncertainty by

$$H[X \mid p_i] = -\sum_i p_i \log p_i$$

bits compared to the case where each value is equally probable.

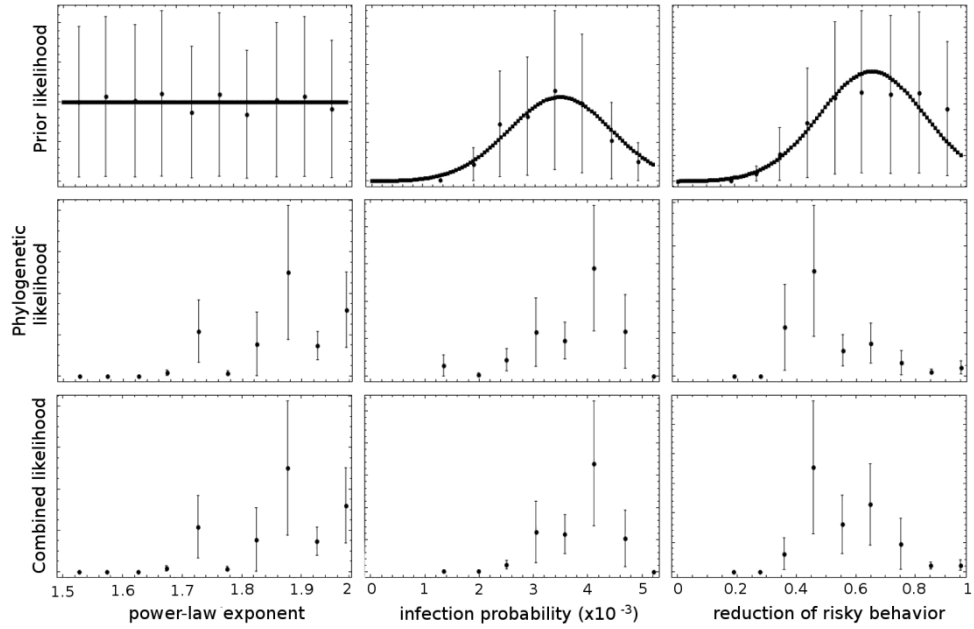

Figure 13: The prior knowledge, additional knowledge from phylogenetic data, and the combined knowledge about three important epidemiological parameters: the network topology, the per-act infection probability, and the risk behavior reduction upon diagnosis. Each range of parameter values was divided into 10 bins in order to calculate the Shannon entropy. For the latter two parameters we used a maximum value equal to the 95% quantile.

Estimating the amount of information about a continuous variable from sampling is problematic. This is because a continuous variable is infinitely precise and requires an infinite number of bits to be determined, which is impossible to gather from a finite number of samples. Fortunately we are

not interested in values of infinite precision for practical purposes. Therefore we divide the range of values of each parameter into 10 bins of equal width, which is sufficient to support our main finding.

Although the factor between the per-act infection probabilities of undiagnosed and diagnosed MSM is an expected 0.75, the mean value in the simulations becomes slightly less because we exclude all parameter sets with a factor higher than 1. Such values would imply that MSM would increase their risky behavior upon diagnosis, which is unlikely and unsupported by the literature. The prior probability distributions in the top row in Figure 13 are therefore fitted to the sets of values used in the simulation. It represents more faithfully the prior knowledge of the simulations and prevents overestimating it.

We list the estimated amounts of prior information, phylogenetic information, and combined information in Figure 13. The prior information is calculated from the probability distribution from literature, which is equiprobable for the network topology and a normal distribution for the per-act infection probability and change of behavior upon diagnosis. The `phylogenetic information' is calculated from the probability distribution estimated from the likelihoods of the simulated cluster-size distributions using Eqs. (3) and (4).

The combined information is calculated from multiplying the prior probability distributions with the phylogenetic probability distributions. This means that if the prior knowledge assigns zero probability to a value whereas the phylogenetic knowledge would assign non-zero probability, the combined knowledge is the zero probability. It also means that the likelihood of a value will be highest if the two sources of knowledge agree, i.e. assign equal probability, and decreases as the difference between the two probabilities increases. As prior knowledge we used the uniform and normal probability distributions, depicted as black lines in the upper figures in Figure 13.

The number of bits of information contained in the knowledge of Figure 13 is calculated using Shannon entropy. For each parameter we show in Figure 14 the remaining uncertainty about the parameter value for the prior knowledge, phylogenetic data, and their combination, respectively. We measure the remaining uncertainty as the fraction of the required information that is still missing despite having certain knowledge. Since we divided each parameter value range into 10 bins, the required information to identify a parameter value is $\log_2 10 \approx 3.32$ bits.

The prior knowledge is agnostic about the power-law exponent of the network topology, indicated by a uniform distribution, so the information it contains is zero. The knowledge provided by the phylogenetic data is not very specific about the exponent, but it tends to exclude the low values. All in all, the knowledge contains an estimated 31% of the necessary information to uniquely identify the power-law exponent. The combined knowledge is identical since there is no prior knowledge.

The phylogenetic data is roughly equally specific about the per-act infection probability. The prior knowledge is a normal distribution, which prefers values close to the mean, and provides 17% of the required information. Although the phylogenetic data has a preference for slightly higher parameter values, it appears to agree with the prior knowledge. The information provided by the phylogenetic data is 32%. Since both sources of knowledge tend to agree, the combined knowledge is even more specific and provides 41% of the required information.

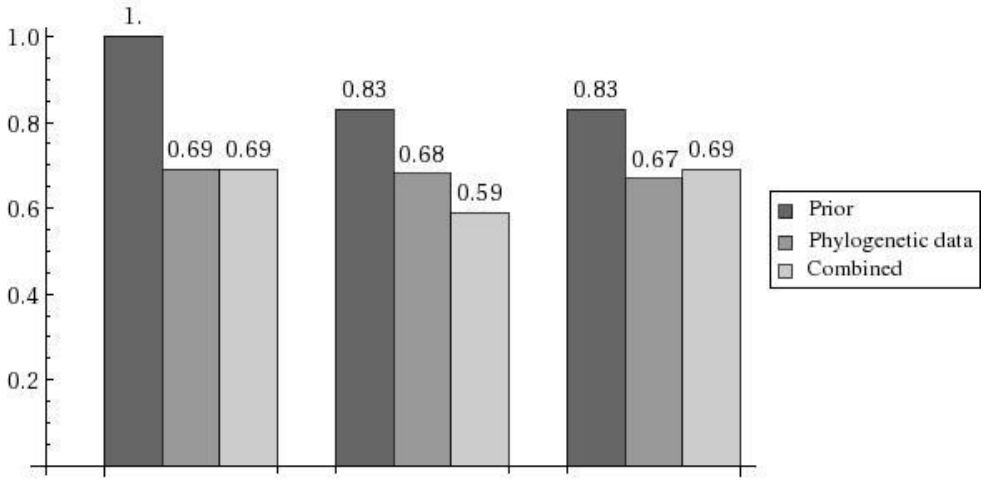

Figure 14: The remaining uncertainty given the prior, phylogenetic, and combined knowledge about the parameter values corresponding to the probability distributions in Figure 13. Here we show the fractions of uncertainty with respect to the maximum $\log_2 10$. The lower the bar, the more information is contained in the corresponding knowledge.

The prior knowledge about the reduction of risky behavior turns out to be identical to that of the per-act infection probability. This is because the reduction of risky behavior is calculated directly from the infection probabilities, so they share the same source of prior uncertainty. The phylogenetic data is also roughly equally informative about this parameter value at 33%. However, the phylogenetic data prefers significantly lower values than the prior knowledge. Since both sources of knowledge disagree, the combined knowledge becomes slightly less specific and decreases to 31%.

These measured amounts of information provided by the phylogenetic data should be used with caution. This information tends to be overestimated by our methodology. This is because for a finite number of simulation runs, the distribution of likelihoods tends to be jagged instead of smooth. A jagged likelihood distribution provides more information than the corresponding smooth distribution because some values become more likely than others, so

it is more specific about the value. The more simulation runs, the smoother the estimated distribution of phylogenetic likelihood, which implies a smaller overestimation of the phylogenetic information about parameter values. Here we used 500 simulation runs due to technical limits, but from Figure 13 we see that the estimated likelihood distributions are not smooth. This is especially true for the phylogenetic information about the power-law exponent. Therefore the absolute amounts of information and their ratio should not be relied upon, however the results are sufficient to answer the question whether phylogenetic data contains information about the epidemiological parameters.

In conclusion, we find indeed that the cluster-size distribution of genotypes of HIV-patients contains significant information about the epidemiological parameters. The amount of information was estimated using 500 simulations of 6000 agents in a dynamical network. Each simulation run was initialized with semi-random parameter values based on the literature, and its predicted cluster-size distribution was compared to the observed distribution based on sequence data from the United Kingdom. This provides a likelihood function of all possible sets of parameter values, and we calculated the corresponding amount of information it provided. We estimated that the cluster-size distribution provides up to one third of the required information about the sexual network topology, the per-act infection probability, and the change of behavior upon diagnosis with HIV. The information contained in the cluster-size distribution is a lower bound of the information contained in phylogenetic data because the cluster-size distribution is a subset of the data. The present work shows that ambiguous and incomplete phylogenetic data indeed contains information about how a disease has been transmitted through a population.