Effective HIV prevention requires knowledge of the structure and dynamics of the social networks across which infections are transmitted. These networks most commonly comprise chains of sexual relationships, but in some populations, sharing of contaminated needles is also an important, or even the main mechanism that connects people in the network. Whereas network data have long been collected during survey interviews, new data sources have become increasingly common in recent years, because of advances in molecular biology and the use of partner notification services in HIV prevention and treatment programmes. We review current and emerging methods for collecting HIV-related network data, as well as modelling frameworks commonly used to infer network parameters and map potential HIV transmission pathways within the network. We discuss the relative strengths and weaknesses of existing methods and models, and we propose a research agenda for advancing network analysis in HIV epidemiology. We make the case for a combination approach that integrates multiple data sources into a coherent statistical framework.

Keywords: contact tracing, epidemiology, HIV, mathematical models, partner notification, phylogenetics, sexual networks, survey data

Introduction

HIV is transmitted through social networks that are formed primarily by the sexual relationships and needle-sharing practices individuals engage in. These networks are important determinants of the magnitude of HIV epidemics [1–6], and the effectiveness of HIV prevention [7–12]. Unfortunately, investigating these networks empirically is difficult, because it requires not only collecting data on sensitive risk behaviours, but also the mapping of relationships that connect people at risk of transmitting and acquiring HIV [13].

In this study, we review how the disciplines of the social sciences, molecular biology, and public health have developed various approaches to collect and analyse network data for the purpose of understanding HIV epidemiology and enhancing HIV prevention and treatment. First, we consider key aspects of network analysis relevant for HIV epidemiology. Next, we provide an overview of existing methods for collecting and analysing network data, and we discuss their relative strengths and weaknesses. Lastly, we emphasize the need for an interdisciplinary approach to network analysis in HIV epidemiology and provide examples of the
synergistic benefits of a combination approach to the challenge of inferring the HIV transmission pathways from which HIV epidemics emerge.

**Network epidemiology and its challenges**

Social networks are sets of links (also called ‘connections’, ‘ties’ or ‘edges’) between individuals (also called ‘nodes’ or ‘vertices’). In common visualizations of networks (Fig. 1), individuals are represented by symbols. The shape and colour of each symbol may capture characteristics relevant to HIV transmission (e.g. gender, race, ethnicity and viral load). Lines that connect the symbols represent links, and the line width and dotting pattern can reflect link characteristics (e.g. frequency of sexual intercourse, condom use, needle sharing). A large number of mathematical tools exist to analyse the data that underlie networks [14]. These tools allow to, for example, identify who is connected to whom, or measure the (social) distance between people living with HIV (PLWH) and uninfected people.

Network analysis in HIV epidemiology revolves around identifying HIV transmission pathways, that is, the subsets of links and individuals across which HIV can spread. Some of these pathways are realized; they connect the PLWH that transmitted HIV to one another. Other pathways represent chains of potential transmission events; they link PLWH to individuals who are not (yet) infected, but are at risk of acquiring HIV in the future because of their network connections. Being able to map and anticipate these pathways helps determining which individuals are most at risk of HIV infection, and select the interventions most likely to interrupt HIV transmission.

The social science approach

The workhorse of the social science approach is the egocentric network survey [13]. In this type of study, a random sample of individuals from the population of interest (Fig. 1c) is asked to provide information about their recent sexual and/or needle sharing partners (e.g. their age, race, and gender), and to describe the characteristics of these relationships (e.g. start and end dates, condom use). Respondents may also be invited to test for HIV infection [24].

Egocentric surveys have been conducted worldwide [25–34]. They permit measuring characteristics of the personal networks of respondents such as homophily (the propensity to engage in partnerships with others who share similar characteristics), or concurrency (the likelihood of having more than one ongoing relationship at one point in time) [35]. But they do not provide data on HIV transmission chains because the partners of sampled respondents are not typically enrolled in the study [13].

In this data context, potential HIV transmission pathways can only be inferred. Network inferences have greatly improved recently, due to the development of exponential random graph models (ERGMs). ERGMs are a family of statistical models that can accommodate the interdependencies between individuals that characterize network datasets [36–42]. They were originally developed for the analysis of complete network datasets in which all individuals and links are listed. But they can also be used with incomplete data from egocentric studies under certain simplifying assumptions [43–45].
Fig. 1. Graphical representations of simulated social networks and HIV transmission pathways. All data from Fig. 1 are synthetic and were generated using Simpact, a freely available agent-based modelling tool for simulating HIV transmission in dynamic sexual networks [23]. (a) Complete (cumulative) network of all relationships that were formed over a 30-year time period between members of the simulated population. Thicker links represent longer relationships. Histograms at the top of (a) represent the degree distribution (lifetime number of partners) in the population, the distribution of relationship durations, and the distribution of log_{10} SPVL among PLWH. (b) Each graph represents sexual relationships that existed in the first, middle, and last 10-year time slice. (c) Simulated egocentric network survey conducted in the population represented in (a). Larger symbols are participants in the survey and the black edges represent their reported relationships in the last 10 years prior to the survey at the end of the 30-year simulation period. (d) In grey, the potential transmission pathways resulting from the introduction of HIV into the simulated population from the index case (framed square). The black links represent the realized HIV transmission pathway. (e) Phylogenetic tree, reconstructed from virus samples. (f) Simulated partner notification investigation starting from the framed circle, that is, the first person to present to a clinic and be diagnosed with HIV in this simulated population. Arrows indicate relationships reported during partner notification interviews. Black arrows connect individuals that were seropositive after testing. PLWH, people living with HIV; SPVL, set-point viral load.
To infer potential HIV transmission pathways from an egocentric study, information about the process of network formation (who forms relationships with whom) first needs to be extracted from the survey data. To do this, an ERGM is fitted to the reported network data to determine the extent and patterns of homophily or concurrency in the population [44]. Next, a set of complete networks is generated on the basis of this model of network formation, and the size and composition of the population [39,46]. Finally, HIV epidemics are simulated on the generated complete networks, by infecting a small number of individuals at random and recording who subsequently gets infected and when. This approach has provided key insights into HIV transmission dynamics [47–50]. Recent advances in ERGM methodology include the development of tools for modelling dynamic networks [51], and freely available software [52,53].

Apart from ERGMs, agent-based models have also been used to infer network characteristics from egocentric data [9,54–58]. In these models, the social network and the subsequent spread of HIV emerge from behavioural and biological 'rules' specified by the analyst at the level of the individual [59].

Naturally, the accuracy of inferences about HIV transmission pathways derived from egocentric data depends on the validity of the model of network formation. If salient aspects of partner choice are not included in this model (either because they are left out by the analyst, or because they were not documented during the survey), then the resulting network inferences may be affected. For example, in a population where PLWH primarily seek to form new partnerships with other PLWH [60–63], ERGMs that do not account for serosorting may misrepresent patterns of exposure to HIV in the population.

Furthermore, important groups of individuals such as mobile and marginalized key population (sex workers, injecting drug users) may be underrepresented in egocentric surveys [64]. Among included individuals, response rates in egocentric surveys are often imperfect and may be associated with HIV transmission risks. For example, PLWH who are aware of their status are significantly less likely to participate in surveys that include HIV testing [65].

Finally, egocentric data provide incomplete and often inaccurate data on the links that connect individuals in a population. To minimize respondent fatigue and recall bias, survey respondents often omit to report some of their relationships during interviews [66–70] and misreport the characteristics of some of their partners [71,72]. These various forms of missing data affect network inferences from egocentric survey data [73].

Social scientists have also occasionally attempted to map the complete network of a population [5,18,74,75]. In such sociocentric studies (also called 'network censuses') [13], each network member is interviewed and asked to provide the names of their sexual or injection partners. These nominations are then linked to identify the relationships that connect population members [76]. Potential and realized HIV transmission pathways can be observed directly if the study also includes HIV testing [5,74]. Sociocentric studies have been restricted to small, isolated population because of the challenges of delineating network boundaries, and linking survey data to identify relationships [13,77]. In addition, as during an egocentric study, some network members may be absent, decline to be interviewed, or omit to report some of their relations. Sociocentric studies are thus never fully complete.

### The molecular biology approach

Whereas social scientists often start from a random sample of individuals irrespective of their HIV status, the molecular biology approach focuses on PLWH. For HIV, as for many other retroviruses, the rate at which viral populations undergo genetic changes within each HIV-positive person (or 'host') is orders of magnitude faster than the rate at which they are transmitted between hosts [78]. These genetic differences between viral populations in different hosts can be used to infer the most likely evolutionary history of the pathogen. The molecular study of networks then entails obtaining viral sequences from PLWH [79], and grouping HIV sequences by genetic similarity [80]. These groupings are depicted as phylogenetic trees (or phylogenies), where tips in the tree represent PLWH, the branching pattern indicates the genetic similarity between sequences from different PLWH, and the internal nodes of the tree represent past transmission events (Fig. 1e). The phylogenetic tree is thus a proxy of a realized HIV transmission pathway.

Phylogenies have been used to identify HIV transmission clusters, that is, groups of PLWH with highly similar viral populations and who are likely connected by an HIV transmission pathway [81–88]. When cluster analysis incorporates the demographic, behavioural, and clinical characteristics of cluster members [89,90], this information may help guide the targeting of HIV prevention and treatment programmes [91]. Viral linkage analysis in HIV clinical trials with serodiscordant couples has enabled more accurate estimation of the efficacy of early antiretroviral therapy (ART) [92] and genital herpes...
suppression [93] to prevent HIV transmission. It has also been used to substantiate [94,95] or reject [84,96] alleged (iatrogenic) HIV infection in forensic medicine, and to discern whether new infections originate from within or outside of the intervention communities during cluster-randomized trials [97–101].

The shape of a phylogenetic tree (also called its ‘topology’) can help elucidate aspects of the broader structure of the social networks and HIV transmission pathways within them. For example, measures of tree asymmetry have been used to test whether individuals form partnerships at random, or rather according to more structured processes. Highly asymmetric trees [102,103], in which some branches led to many infections (Fig. 2c), thus suggest that the underlying social networks may be formed through preferential attachment, leading to scale-free networks [104,105] (with super-spreaders), whereas more balanced trees are associated with random or small-world networks (high interconnectedness among local cliques) [106]. However, different transmission networks can yield phylogenetic trees with similar topologies. Additional summary tree statistics, such as branch lengths, tree width, tree depth, and the occurrence of ‘cherries’ and ‘ladders’ (Fig. 2c), are then needed to differentiate between homogeneous, chain-like and super-spreader transmission networks (Fig. 2b) [106–108]. So-called phylodynamic methods provide an alternative approach that avoids having to simulate complete network data. By combining a particular epidemiological model (e.g. a Susceptible-Infected-Removed model) with phylogenetic data, they permit directly estimating properties of transmission pathways such as the fraction of super-spreaders [109,110].

A major advantage of phylogenetic analyses over interview-based methods is that they are not subject to recall or social desirability bias. And unlike egocentric studies, they elicit indirect connections between individuals who may be at two or more degrees of separation. Molecular biology thus allows longer range investigations of the connectivity of networks, including across geographical regions [88,111,112], age and racial/ethnic groups [113], and between subpopulations [114–117].

There are several limitations associated with the molecular biology approach to network inferences, however. First, the evolutionary model for how genetic changes accumulate over time within and between PLWH typically makes several simplifying assumptions. Frequent model assumptions include neutral evolution (no effects of mutations on viral fitness); no selection of viral subpopulations during transmission events; viral evolution driven by mutation only (not by recombination); and no further transmission after HIV diagnosis. Second, a specific phylogenetic tree is only an ‘estimate’ of the actual true evolutionary history. Reconstruction methods that account for the uncertainty in the tree exist, but are very computationally intensive, thereby limiting the size of datasets that can be analysed. Third, the network inferences drawn from phylogenetic data depend on the sampling density, which is the proportion of PLWH for whom a viral sequence is available [118]. If it is too low, then it will be difficult to link PLWH to the source of their infection, and the phylogenetic tree will be sparser than it really is [79,88]. A recent study in Botswana suggested that a sampling density of 50–70% was required for accurate identification of transmission clusters [111,118]. This is particularly problematic for populations with high HIV prevalence and incidence, where a prohibitively large number of HIV sequences may be needed to achieve an appropriately high sampling density. Fourth, further sampling biases emerge when the individuals for whom sequences are available are not a random sample of the population of PLWH [100,119]. A large number of phylogenetic studies, see [120–122] for recent examples, rely on sequences routinely collected for antiretroviral resistance testing among PLWH who are about to initiate treatment. Phylogenetic trees

Fig. 2. The complete network influences the realized transmission pathway and subsequent phylogenetic tree, but only to a certain degree. (a) Two examples of differing complete networks. (b) Many different HIV transmission pathways may emerge following the infection of a single individual (framed circle). Individual characteristics such as sex (circles/squares) can restrict the possible transmission networks, but mixed modes of transmission (e.g. heterosexual, MSM, and needle-sharing) further extend the range of possible transmission networks. (c) The transmission networks from (b) leave a trace in the phylogenetic tree topology. The topology of a phylogenetic tree is defined as the structure of the tree with its tips (or ‘leaves’, i.e. the end points of the tree), without paying attention to branch lengths and left-right ordering of branches and leaves. Topologies can be compared using imbalance (topologies 2 and 3 are less balanced, i.e. more asymmetrical than topology 1), but different transmission networks (Ib and IIa) may result in the same topology (topology 2). In these cases additional information such as branch lengths and the number of cherries and ladders may help differentiate between transmission networks.
Partner notification has several other strengths. First, similar to phylogenetic studies, it enables routine collection of network data, integrated within HIV care and treatment programmes. Second, partner notification enables tracing of networks well beyond the narrow geographical boundaries of an administrative unit. In the early days of the AIDS epidemic in the United States, for example, contacts of the so-called ‘patient zero’ were traced to New York, Los Angeles, and even abroad [133]. Third, partner notification may help uncover subsets of HIV transmission pathways that often remain hidden in egocentric or phylogenetic studies. These include hard-to-reach high-risk groups [134], stigmatized population that rarely attend health facilities and mobile individuals who are less likely to be included in traditional survey sampling frames. Partner notification is more effective at reaching such groups because its network sampling process is ‘adaptive’ in the sense that it uses information provided directly by network members to guide the selection and recruitment of individuals [43].

Partner notification also has limitations. Whereas partner notification data indicate who notified whom about HIV infection (Fig. 1f), it is difficult to determine who infected whom from such data. In some relations, the person who acquired the infection may in fact seek care – and thus be offered partner notification services – before the transmitter. Additional data such as the date of potential exposure to HIV, and are invited to visit a health facility for HIV testing and linkage to care, if indicated. The process of contact solicitation, contact tracing and testing is repeated for notified partners who test positive for HIV, but it stops for partners who are HIV-negative. Several electronic platforms have emerged that facilitate this process [125–127]. Although face-to-face partner notification is often preferred by index cases, e-postcards, or mobile text messages are also increasingly used [128,129].
A combination approach for network analysis in HIV epidemiology

All three approaches described in previous sections shed light on key aspects of the social networks that propagate HIV, but from different angles. In combination, they may help uncover previously unknown features of HIV transmission pathways. We briefly illustrate how an interdisciplinary approach to network epidemiology could help improve our ability to understand processes of network formation and prevent future HIV infections.

Egocentric network data are affected by underreporting of links. This may bias the results of ERGMs or other models that are solely informed by egocentric network data; the inferred networks and HIV transmission pathways will be less dense than the true networks. If underreporting is especially prevalent among specific population groups (e.g., unmarried women in sub-Saharan population) [67,138], then these groups may appear at decreased risk of HIV acquisition, when, in fact, their level of exposure may be substantial. But if HIV sequence data are available for the population of interest, phylogenetic analyses could help adjust for such biases in network models. The topology of transmission clusters, for example, could be used to estimate the mean and variance of the distribution of the numbers of partners individuals have, possibly stratified by gender. These parameters can then be used as correction factors in the network model.

On the other hand, it may also be difficult to predict the potential impact of an HIV prevention intervention based solely on molecular data. Phylogenies only document realized HIV transmission pathways, whereas there may exist a (possibly large) number of competing but unrealized transmission pathways (Fig. 2). For instance, while almost half of HIV transmissions among MSM in the United States occur when the index case is in their first year of HIV infection [139], there may be other potential transmission pathways that involve chronically infected PLWH. Therefore, reducing the transmission potential during the early stages of infection could have a smaller impact on HIV incidence than expected if individuals are connected by more than one potential transmission pathway (Fig. 1d). Such network redundancies are unobserved in the molecular data, but can be recorded directly through sociocentric studies or during partner notification [5,140,141], and can be inferred from egocentric data using ERGMs [48].

As a final example, linking HIV drug resistance data (from HIV sequence analysis) to indicators of ART adherence as well as behavioural, relational, and sociodemographic details of PLWH could help identify subgroups in need of intensified treatment monitoring and support. Superimposing this information onto reconstructed phylogenies and inferred transmission networks may also help predict the evolution of transmitted and secondary drug resistance and identify targetable correlates of residual transmission by PLWH on ART.

Klovdalh et al. [142] successfully pioneered integration of multiple network data sources in an epidemiological study of tuberculosis, and an increasing number of studies in HIV epidemiology now also combine multiple sources of data on social networks [88,143–151]. The recently launched Phylogenetics and Networks for Generalized HIV Epidemics in Africa consortium (PANGEA-HIV) aims to sequence 200,000 total HIV genomes from Uganda, Botswana, South Africa, and Zambia, and link sequence data to clinical, demographic, and behavioural data [152].

There are, however, considerable methodological challenges to a combination approach to network analysis in HIV epidemiology. Consider the small population of 10 individuals in Fig. 4, in which all three sources of data described above are available (Fig. 4b–d). In this setting, combining data sources can yield a significantly more complete picture of the social networks, and the location of PLWH within these networks (Fig. 4c). But this still requires a network inference step to identify networks that are compatible with the reported egocentric, partner notification, and phylogenetic data.
Unfortunately, network inference using multiple data sources is complicated by a number of factors. First, contrary to egocentric data, partner notification, and phylogenies are not typically based on random samples of the population. As a result, the probabilities of inclusion of each individual and link in the network dataset are unknown, and this may affect the results of ERGMs and other existing models of network inference. Second, in some settings, the various network datasets may not be linked at the individual and/or relationship levels. We would not know whether an individual who reported three links during an egocentric survey is the same person who tested HIV-positive or refuses to name their partners. Partner notification stops if the partner of a PLWH is HIV-negative or refuses to name their partners. (c) Public health officials conduct partner notification. Interviews are conducted with PLWH who present at health facilities (‘+’ in thick borders). Phylogenetic analysis of HIV sequence data to identify longer range connections between new and previously interviewed individuals, or yet unidentified intermediate transmitters (empty octagon). (e) Ultimately, a more complete network picture emerges from the union of these three data collection approaches, through network modelling. PLWH, people living with HIV.

**Fig. 4.** Approaches from social science, public health and molecular biology can synergistically improve inferences about social networks and HIV transmission pathways. We schematize the case of a population where network data have been collected through the three data collection approaches described above. (a) The starting point is a population whose composition and network connectivity is unknown. The goal is to infer the potential transmission pathways through which HIV can spread within this population. (b) The social science approach conducts egocentric studies, which draw on a random sample of men and women (blue and orange symbols). These respondents report the extent and composition of their personal networks. (c) Public health officials conduct partner notification. Interviews are conducted with PLWH who present at health facilities (‘+’ in thick borders). Partner notification stops if the partner of a PLWH is HIV-negative or refuses to name their partners. (d) Phylogenetic analysis of HIV sequence data to identify longer range connections between new and previously interviewed individuals, or yet unidentified intermediate transmitters (empty octagon). (e) Ultimately, a more complete network picture emerges from the union of these three data collection approaches, through network modelling. PLWH, people living with HIV.

Unfortunately, network inference using multiple data sources is complicated by a number of factors. First, contrary to egocentric data, partner notification, and phylogenies are not typically based on random samples of the population. As a result, the probabilities of inclusion of each individual and link in the network dataset are unknown, and this may affect the results of ERGMs and other existing models of network inference. Second, in some settings, the various network datasets may not be linked at the individual and/or relationship levels. We would not know whether an individual who reported three links during an egocentric survey is the same person who tested HIV-positive or refuses to name their partners. Partner notification stops if the partner of a PLWH is HIV-negative or refuses to name their partners. (c) Public health officials conduct partner notification. Interviews are conducted with PLWH who present at health facilities (‘+’ in thick borders). Phylogenetic analysis of HIV sequence data to identify longer range connections between new and previously interviewed individuals, or yet unidentified intermediate transmitters (empty octagon). (e) Ultimately, a more complete network picture emerges from the union of these three data collection approaches, through network modelling. PLWH, people living with HIV.

**Conclusion**

Data on sexual networks come from an increasingly diverse array of sources, but each of these sources only document parts of the networks through which HIV may spread. Egocentric network surveys suffer from non-response, social desirability bias, and the inability to probe beyond the immediate network connections of individuals. Through partner notification services, realized, and potential HIV transmission pathways may be partially revealed, but in resource-poor settings with generalized HIV epidemics offering partner notification services may require prohibitively large investments. Phylogenetic tree analysis permits reconstructing parts of the HIV transmission chains by linking genetically related infections, but to be informative, HIV sequence data must be available for what may be an unfeasibly large sample of PLWH. Novel methods to combine these data sources are beginning to emerge from the collaborative efforts
of experts in computational biology, social science, statistics, public health, and epidemiological modelling. Further advances in network analysis for HIV epidemiology will require important methodological developments in network modelling, as well as a long-term, global commitment from researchers and funding agencies to ensure open access to analytical tools and multifaceted network datasets that include HIV sequences along with behavioural, demographic, clinical, and programmatic information.

Acknowledgements

We thank Jori Liesenborgs, who developed and documented the Simpact Cyan program (C++), as well as the RSimpactCyan front end (R package), and Gavin Hitchcock, Roxanne Beauchard and Niel Hens for their helpful comments on an earlier version of this study. We acknowledge support from the Eunice Kennedy Shriver National institute of Child Health and Human Development, grant # R03HD071122 (PI: Helleringer).

Conflicts of interest

There are no conflicts of interest.

References

Network data and models in HIV epidemiology Delva et al. 2019


