Journal of Theoretical Biology 613 (2025) 112218

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Large connected components in sexual networks and their role in HIV transmission in Sub-Saharan Africa: A model-based analysis of HPTN 071 (PopART) data☆

Francesco Di Lauro [a,*], William J.M. Probert [a], Michael Pickles [b], Anne Cori [b], Robert Hinch [a], Luca Ferretti [a], Jasmina Panovska-Griffiths [a,h], Lucie Abeler-Dörner [a], Rory Dunbar [c], Peter Bock [c], Deborah J. Donnell [d], Helen Ayles [e,f], Sarah Fidler [g], Richard Hayes [f], Christophe Fraser [a], the PANGEA Consortium

[a] Pandemic Sciences Institute, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[b] School of Public Health, Imperial College, London, United Kingdom
[h] The Queen's College, University of Oxford, Oxford, United Kingdom
[c] Desmond Tutu TB Centre, Department of Paediatrics and Child Health, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[d] Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[e] Zambart, Lusaka, Zambia
[f] London School of Hygiene and Tropical Medicine, London, United Kingdom
[g] Department of Infectious Disease, Imperial College London, London, United Kingdom

ARTICLE INFO

Keywords:
Infectious disease model
HIV model
PopART trial
Sexual networks

ABSTRACT

The HIV epidemic in sub-Saharan Africa is historically characterised by high levels of prevalence and incidence. With the global effort to reach UNAIDS 95-95-95 targets, the scaling-up of HIV treatment, and focused preventive interventions, incidence has been declining over the past decade, albeit non-consistently across different sex and age groups. Two questions remain to be addressed to help tailor setting-specific interventions and allocate resources optimally. Firstly, are there unidentified demographic groups that are sources of transmission? Secondly, what are the patterns of decline in incidence across different groups? Model-based assessment is a valuable tool for the design of focused interventions and to answer these questions. PopART-IBM, an individual-based model calibrated to (anonymised) age-and-sex stratified data, was developed in the context of the HPTN-071 (PopART) trial, and it offers a unique opportunity to explore such questions in the context of high-burden HIV communities in Zambia and South Africa. The outputs of the model include the full HIV transmission and partnership networks. In this work, we explore these and show that the sexual partnership network exhibits a large connected component, usually comprising over 40 % of the population, in each of the studied communities. An analysis of the large connected component reveals that it is formed by young people (20–40 years old) and is centered around the most sexually active individuals of the community. At the same time, many individuals in the large connected component only have one partner, highlighting the complex dynamics of risk correlations in a population. Inspecting the transmission network reveals that, on average, more than 80% of transmissions occur among individuals belonging to the large connected component. These findings indicate that populations consisting of young and highly sexually active individuals should be given high priority when designing or deploying interventions.

☆ This article is part of a special issue entitled: 'Novel methods and lessons learned from COVID-19' published in Journal of Theoretical Biology.

* Corresponding author.

https://doi.org/10.1016/j.jtbi.2025.112218
Received 31 October 2024; Received in revised form 7 July 2025; Accepted 14 July 2025
Available online 16 July 2025

1. Introduction

The contact structure of a population is important in determining how epidemics spread (Helleringer and Kohler 2007; Read and Keeling 2003; Keeling 2005) and plays a fundamental role in shaping transmission networks. This is particularly true for sexually transmitted infections, as only specific types of contacts can lead to transmission. However, gathering detailed data on human contact networks is a difficult task, as detailed sexual practices and behaviours are often not fully shared in surveys, although increasing amounts of data are available through monitoring surveys (Justman, Mugurungi, and El-Sadr 2018), trials (Perriat et al. 2018), and cohort studies (Tanser et al. 2008). Inferring detailed and structural properties of the contact network remains challenging (Schaub, Segarra, and Tsitsiklis 2020). In the context of HIV, for instance, we observe two very different epidemic patterns when we look at outbreaks in Europe and in Sub-Saharan Africa using phylogenetic methods (Magosi et al., 2022). While the former is characterised by heterogeneous transmission and large cluster sizes, i.e. with densely connected phylogenetic sequences, the latter is characterised by smaller, more homogeneous clusters, where many sequences are linked to no, or very few, onward transmissions. Direct measurements of these dynamics remain difficult, as only a fraction of transmissions can be inferred from available data. Therefore, the presence of core groups, that is, subpopulations who can drive the epidemic, cannot be excluded by phylogenetic transmission clusters analysis or based on survey reports.

In this context, modelling can be a valuable tool to explore whether hypotheses about the transmission network are consistent with the available data. In HIV epidemiology, models have been used extensively to provide insights into the mechanisms of HIV transmission, be they biological (Bellan et al., 2015) or behavioural (Watts and May 1992; Hallett et al. 2007). The recent increase in the amount of data available through surveillance, trials, and cohort studies allowed the development of increasingly complex models. The PopART-IBM model is calibrated to age-and-sex stratified data from multiple sources on demographics, HIV prevalence, awareness of HIV status, ART status, and viral suppression for the HPTN 071 (PopART) trial in Zambia and South Africa (Pickles et al. 2021, Hayes et al. 2019, Probert et al., 2022), a large cluster-randomised trial of universal testing and treatment, with 21 communities participating in the study. An interesting feature of the model is that it generates the partnership network dynamics as an emergent property from few assumptions related to risk assortativity, rates of formation/dissolution of partnerships, and individual sexual activity, which in turn are inferred from survey data from a longitudinal HIV incidence cohort, named the Population Cohort, comprising roughly 2000 individuals per community randomly sampled from the adult population, who were interviewed annually for four years (Hayes et al. 2014).

The model outputs include the full partnership history of everyone in the simulated population for each of the 21 communities that were part of the trial, as well as the transmission networks. This allows us to investigate contact-structure properties which play a role in determining the epidemic dynamics. In this context, a network of sexual partnerships can be defined at each timestep, as the network of sexual relations that are ongoing at that timestep among sexually active individuals. This network is dynamic, as partnerships form and dissolve over time, and it is formed by individuals (nodes) and (temporal) partnerships between individuals (links). When the model is fitted to data, the partnership network has a large connected component (lcc) across all the calibrated runs (Fig. 1). The size of the large connected component varies between 40 % and 80 % of the sexually active population across all simulation runs, as depicted in Fig. 2. An lcc is a subset of the network in which any pair of nodes is connected by at least one unique path, whose size is non-negligible with respect to the size of the network. This means that for any two people belonging to the same component, there is a chain of ongoing sexual relationships between them. A network with an lcc may potentially accommodate long transmission chains and therefore sustain outbreaks more easily than a sparser contact network. A component greatly affects the risk of acquiring/transmitting a pathogen for individuals who are part of it (Robinson, Everett, and Christley 2007; Kiss, Green, and Kao 2006). In the literature, there is some evidence of the existence of lcc in sexual networks in Africa related to the spread of HIV (Helleringer and Kohler 2007). From a theoretical point of view, it had been studied (Boily et al., 2002; Di Lauro et al., 2020) how static networks with similar macroscopic properties could generate a variety of contact structures, which would however experience different disease dynamics, highlighting the interplay between transmission patterns and contact structures. Studies disagree to which extent concurrent relationships play a role. For example, Tanser et al. 2011 do not find a strong association between concurrent partnerships and risk of HIV acquisition, while others find the opposite (Hudson 1996; Vijver et al., 2013).

In this paper, we aim to use the information gathered during the HPTN071 (POPART) trial, enriched with modelling to characterise the role of large connected components in the HIV epidemic dynamics in Zambia and South Africa over the period 2010–2019. To achieve this aim, we have two objectives. Firstly, we describe the demographic characteristics of the large connected component and identify the core

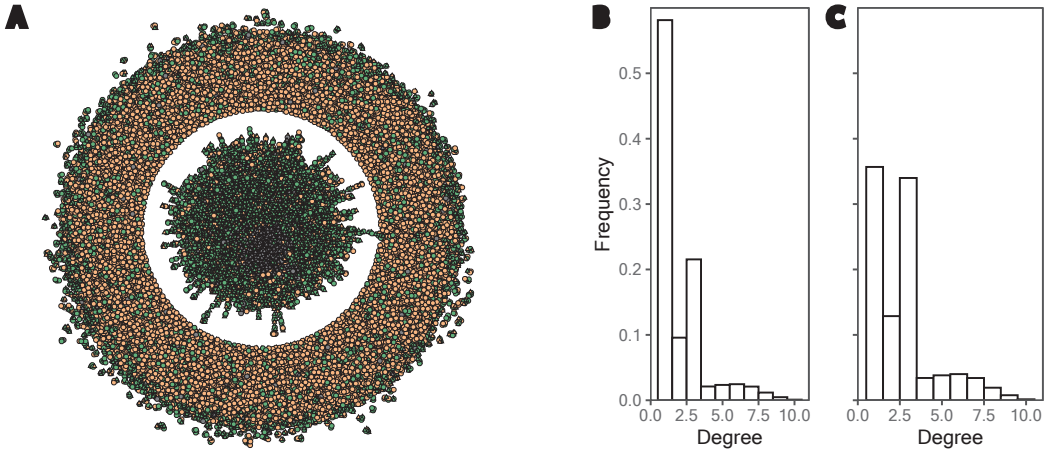

Fig. 1. A – Snapshot of the partnership network from one calibrated run of PopART-IBM (all the ongoing relationship during the first timestep of 2010). Triangles represent males, circles represent women, node colours show sexual activity group of the person (yellow – low, green – medium, black – high). In the centre is the large connected component of the network, whose size in this run is more than half of the total population. B – Degree distribution of the whole network of sexually active people; C – Degree distribution of the large connected component, that is the distribution of the number of partners over one year in the simulated population.

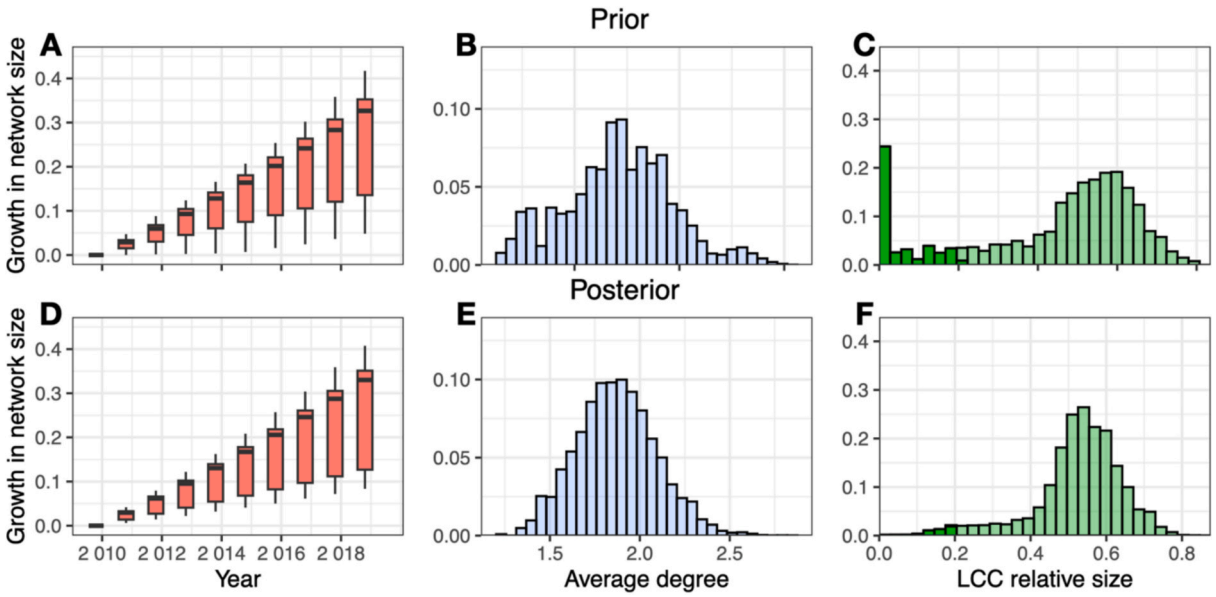

Fig. 2. Partnership network statistics over a 10 year period for all the communities, comparison between priors (top row) and posteriors (bottom row) of the model. In both cases, results are taken from 100 parameter combinations per community. A/D – Relative size growth since 2010 of the number of individuals having at least one partner over ten years, i.e. partnership network size. The absolute network size in 2010 across runs is 59,516 (34971, 73738) B/E – Average degree, i.e. average number of partners per sexually active individual, as of the beginning of 2019. C/D – Relative size of the largest component, i.e. proportion of people in the partnership network that belong to the largest connected component, as of the beginning of 2019. The darker columns indicate the proportion of simulations resulting in largest component size smaller than 20 % of the sexual network.

groups responsible for sustained transmission over the study period. Secondly, we explore the distribution of the number of secondary infections per primary infection – the offspring distribution – to identify how many infections emerge from HIV-positive individuals and to explore the role of large bursts of transmissions.

2. Methods

2.1. HPTN 071 (PopART)

We use modelling to analyse the sexual networks emerging from the HPTN 071 (PopART) trial (Hayes et al. 2014), a cluster-randomised trial conducted in Zambia and South Africa between 2013 and 2018, in 21 communities with a total population of about 1 million. HPTN 071 was a cluster-randomized trial that demonstrated how universal testing and treatment decreased community-level HIV incidence compared to the standard of care. A detailed description of the HPTN 071 trial and its outcomes can be found in (Hayes et al. 2019); in summary, the study was conducted on 21 urban or peri-urban communities, characterised by high HIV prevalence. Communities were divided into seven triplets (4 in Zambia, 3 in South-Africa), matched based on geographical location and HIV prevalence. Communities within each triplet were randomly assigned to one of three arms: arm A, where the full intervention was deployed and individuals were provided ART regardless of their CD4 count, arm B, where universal testing and linkage to care were deployed, while treatment with antiretrovirals followed national guidelines, and arm C, the control group. Universal ART was implemented in both countries in 2016, aligning arms A and B of the trial. The prevention package was delivered to households by means of community HIV care providers (CHiPs). Outcomes were measured in a cohort of approximately 2000 people aged 18–44 randomly sampled from each community, that were followed yearly, named the Population Cohort (PC).

The data sources included the Population Cohort surveys (collected between 2014 and 2018), the data recorded by CHiPs (2013–2017), Demographic Health Surveys (Corsi et al., 2012) in Zambia (2002, 2007, 2013), and Human Sciences Research Council surveys in South Africa (Pickles et al. 2021). Both include age-and-sex stratified data regarding HIV prevalence, proportion of individuals aware of their status, proportion of individuals amongst those aware of their HIV status on ART, and proportion of individuals that are virally suppressed.

2.2. PopART-IBM

PopART-IBM (Pickles et al. 2021) is an open-source, discrete time, agent-based model designed in the context of the HPTN 071 trial, developed to support and help interpret the data collected during the trial. The model broadly consists of six main blocks: (a) a spatial structure, essential to represent sexual interaction between people living in the communities (inside patch) and outside of it (outside patch),(b) a demographic structure which determines how individuals undergo several processes related to demographics, including mortality due to natural causes, (c) a heterosexual partnership formation mechanism that simulates how agents form and dissolve sexual relations, (d) an HIV transmission mechanism to determine the hazard rate at which transmission happens in serodiscordant couples, dependent on multiple health-related factors such as the ART status of the individual, their set point viral load and stage of HIV infection, and whether the male of the couple is circumcised, (e) HIV disease progression, including CD4 decline and HIV-related death, and finally (f) a model for healthcare and prevention, which explicitly simulates ART initiation following national guidelines (or, before 2016, universal ART for all those diagnosed with HIV in arm A), and include adherence to the therapy, as well as the viral suppression process. HIV testing uptake and ART adherence in the model are informed by the data from healthcare facilities and community healthcare providers, disaggregated by age/sex. The model outputs detailed data such as incidence, prevalence, and the full transmission and partnership networks.

The partnership formation/dissolution process, which is the most relevant part of the model for this study, is informed by country-wide health sex surveys (Corsi et al., 2012, Pickles et al., 2021), and by an ad-hoc analysis of the survey data, coming from the baseline survey of the Population Cohort (PC0, 2014), during which more than 38,000 participants were asked questions about their recent partners, such as the number of unique sexual partners they had during the last year and

over their lifetime, whether their partner was from the same community, and about their awareness of HIV, including clinical interventions, such as voluntary medical male circumcision, or non-clinical interventions such as how often they used condoms. A subset of the cohort was given an extended survey, asking more detailed questions about their sexual partnerships (up to three) during the past year. These questions enquired about the participant's partners' age, sex, and behaviours related to risk of HIV acquisition (such as using drugs or alcohol before or during sex), when the relationship begun, frequency of sexual acts, and whether they were aware of their partners' HIV status. The questionnaire also asked whether their partner was long-term (husband/wife), casual, or one time. The results of the analysis of the detailed survey allow defining an age-mixing matrix which governs partnership formation in the model. Further, the model assigns individuals a lifelong sexual activity level, which defines the propensity of forming new relationships. In PopART-IBM, three broad categories of sexual activity are defined based on threshold numbers of lifetime partners (with the threshold being age-dependent) inferred form the survey analysis, that is, "LOW", "MEDIUM", or "HIGH". These define the frequency of sex acts, the average duration of a partnership, that is assumed to be exponentially distributed with parameters that fit the available data by age and sexual activity, and the maximum number of concurrent partnerships that individuals may have at any given time, i.e. up to 1 for LOW sexual activity level, up to 3 for MEDIUM sexual activity, and up to 10 for HIGH sexual activity (see the Supplementary material, Fig. 1). From the PC0 data, it can be inferred that approximately 50% of the population is in the LOW category, 35% in the MEDIUM sexual activity level, and 15% in the HIGH sexual activity category (see the Supplementary material of (Pickles et al. 2021)). PopART-IBM introduces an overall parameter, named assortativity, which describes to what extent individuals in the same risk category tend to prefer relationships with individuals from the same risk group. To allow some flexibility, there is an overall multiplier to the rate of partnership formation, which is equivalent to including in the model a factor accounting for misreporting (1.0 meaning that people on average reported the correct number of partners, greater than 1.0 means that there is underreporting). Both assortativity and misreporting are inferred during calibration, as the data collected do not inform them directly. After calibration, the model suggests that the population structure is assortative by sexual activity group (assortativity between 0.55 and 0.75 in 95 % of the calibrated runs) and that the average underreporting is about 50 %. Finally, a proportion of partnerships are formed across patches, i.e. between individuals from the study communities (referred to as inside patches) and the surrounding areas (known as the outside patch), to simulate partnerships from people who live in different communities. The model does not account for occupation, marital status or other socioeconomic variables, or geographical dispersion (for the outside patch), which could potentially change the partnership formation/dissolution dynamics. The model does not distinguish among different types of partnerships, such as casual or short-term partnerships, albeit duration of a partnership is explicitly simulated. The sexual level assortativity matrix diagonal is allowed to vary, meaning that we allow the model to explore scenarios where people preferentially form partnerships with similar individuals (in terms of sexual activity) or not. Since it is not granted that the number of new partnerships sought by men corresponds to the number of new partnerships sought by women, results of the analysis are adjusted so that the actual number of partnerships sought by men is equal to the number of partnerships sought by women. This was achieved by introducing a compromise parameter which assumes that half the gap is due to underreporting and half due to overreporting. Once a partnership is formed, its duration is distributed according to a Gamma distribution of known shape (depending on age and sexual activity of the individuals involved) and flexible scale, which is fitted by the model. This choice is made to allow some flexibility regarding the mean duration of the partnership while ensuring that results from PC0 analyses are consistent with the results from the model. Of course, if one of the partners dies during the partnership, the partnership is dissolved prior to its end date as computed by the model.

2.3. Partnership network

The discrete-time nature of PopART-IBM allows us to observe the simulated network of active partnerships in each community in its full evolution. The nodes of the network represent individuals who are currently sexually active (defined as having at least one partner), and an (undirected) link between individuals indicates that they form a couple, see Fig. 1. Since we model only heterosexual couples, the network is by design bipartite. The average degree of the network is the average number of partnerships individuals are engaged in at any given time. Since people who are not sexually active do not enter the partnership network, the average degree is always greater than or equal to one, meaning that on average sexually active individuals have more than one partner. It is worth noting that the partnership network is an emergent feature of the model, and, through calibration, of the real-world data we input. This means that the relationship network is not inferred directly from data but rather deduced from the interplay between all the available data plus the assumptions of the model.

In this study, we investigate the presence of lccs in partnership networks emerging from calibrated runs of PopART-IBM, and their role in epidemic dynamics. In network theory, specifically for random networks, when the largest connected component of a network has a size that is comparable with that of the whole network, it is often termed a giant component (Newman 2018). All other connected components typically are of much smaller size, so it is often the case that only one large connected component may be present in the network. In calibrated runs, largest connected components usually include more than 20% of the whole sexually active population.

2.4. Calibration

Although the model needs 350 parameters to run, most are inferred from data or the literature, while 17 are chosen to be calibrated (see the Supplementary material, Table 1). PopART-IBM is calibrated in a Bayesian framework. For each parameter that needs to be inferred, a prior distribution is defined, see the Supplementary material in (Pickles et al. 2021). Calibration of the model is made for each community through an adaptive population Monte Carlo Approximate Bayesian Computation algorithm (Lenormand, Jabot, and Deffuant 2013). The data used for calibration include age-and-sex stratified summary statistics for HIV prevalence, awareness of HIV status, ART coverage, and viral suppression among people living with HIV who are on ART, for a grand total of 248 data points. Each community is modelled separately and is independent from all other communities.

2.5. Statistical analysis

For each community, PopART-IBM is run 100 times, sampling from the posterior distribution of the parameters for each community. We aggregate results across all the communities, irrespective of the country of origin (Zambia and South Africa). In the Supplementary material, some of the analysis done here is carried over without aggregating by country, to identify differences between Zambian and South African epidemics. The network dynamics are captured at one-year resolution scale. Looking at the outputs emerging from the prior distributions, we observe the presence of networks whose lcc size is quite small, relative to the population. This suggests that the model structure is, in principle, able to output partnership networks without an lcc, however calibration to data makes it discard almost all combinations of parameters that do not result in a single large connected component, see Fig. 2.

To explore more systematically risk factors associated with either being part of the large connected component or acquiring an HIV infection, we use logistic regression, both multivariate and univariate,

on different covariates, such as sex, age, and sexual activity level. Unless otherwise specified, all confidence bands shown in plots represent 95 % confidence intervals, generated by aggregating results across all the communities and the runs.

3. Results

3.1. Analysis of the partnership network

The partnership network is bipartite (because only heterosexual relationships are modelled), temporal, and increasing in size over the years, following the population growth trends, see Fig. 2. The network has some heterogeneity, as individuals in it may have any number of concurrent partners from one to ten. The mean degree and its variance vary across different runs. A mean degree greater than one, such as in this case, points towards the existence of a large connected component (Robinson, Everett, and Christley 2007), that is, a connected component containing a significant proportion of the population. During surveys which are "egocentric" and do not allow collection of data beyond the individual level, participants on average reported having more than one current partner, with 80% of relationships formed with individuals from the same community. Further, calibration suggests that on average, the underreporting factor during survey rounds was approximately 1.6 (Probert et al, 2022). These two effects push the average number of concurrent partners estimated from simulations above the critical threshold at which a lcc emerges. The presence of lcc in the network of relationships is relevant from an epidemiological point of view, as in networks that do not exhibit a lccs, transmission chains are typically short-lived.

We compared the prior to the calibrated model outputs and checked if some statistically significant differences emerged. The wide range of parameters in our assumptions allow the model to produce networks with quite small or non-existing connected components. In particular, the South African age structure has relatively more people over 45 than the Zambian one. Since age has an impact on sexual activity in the model (in particular, a decrease in sexual activity level after age 40), the prior distribution in South Africa has a higher peak (Fig. 1B) in the region of parameters that does not allow for a large connected component to emerge than for Zambian communities, see supplementary Figs. 1 and 2 for the same plots disaggregated by country. However, when we calibrate to data, only partnership networks that exhibit a lcc produce model output that is compatible with the epidemic in all the communities, see Fig. 2C. The relative size of the large connected component is not fully identifiable, however more than 75% of the calibrated runs exhibit a large connected component whose size is greater than 50% of the global partnership network, meaning that more than half of the people who are sexually active are part of the large connected component.

3.2. Analysis of the largest connected component

Although PopART-IBM does not currently describe any socio-economic or geographical characteristics of individuals (beyond being either in the community or outside of it), it models demographic and behavioural heterogeneities in individuals. In more than 60% of the calibrated runs, there are more women than men in the large connected component (median ratio of women to men being 1.2), and generally the population is younger and more sexually active than outside of it, see Fig. 3. Although the trends are identical, the main difference between South African and Zambian communities is the age pyramid, which is reflected in the composition of the communities both within and outside of the large connected component (see the Supplementary material).

In terms of the degree centrality, the most connected individuals in the lcc are men with the highest sexual activity levels, whereas comparatively fewer low sexual activity individuals are part of the large connected component. To provide a quantitative analysis, we investigated determinants of being in the large connected component by age/sex/risk/HIV status, by means of both multivariable and univariable logistic regression (Table 1). The most important risk factor associated with belonging to the large connected component is being in the highly active sexual population, followed by being aged 20 to 39. Perhaps surprisingly, being male is associated with a lower probability of belonging to the large connected component (average odds ratio 0.886). The reason is that the network is bipartite, and if men are more sexually active than women, then it is to be expected that more women than men belong to the large connected component. It also reflects the higher percentage of women who are living with HIV in these communities. To summarise, the typical member of the lcc is aged 20–39, and typically more sexually active than the average person.

To show how the lcc impacts epidemic dynamics, we look at trans-

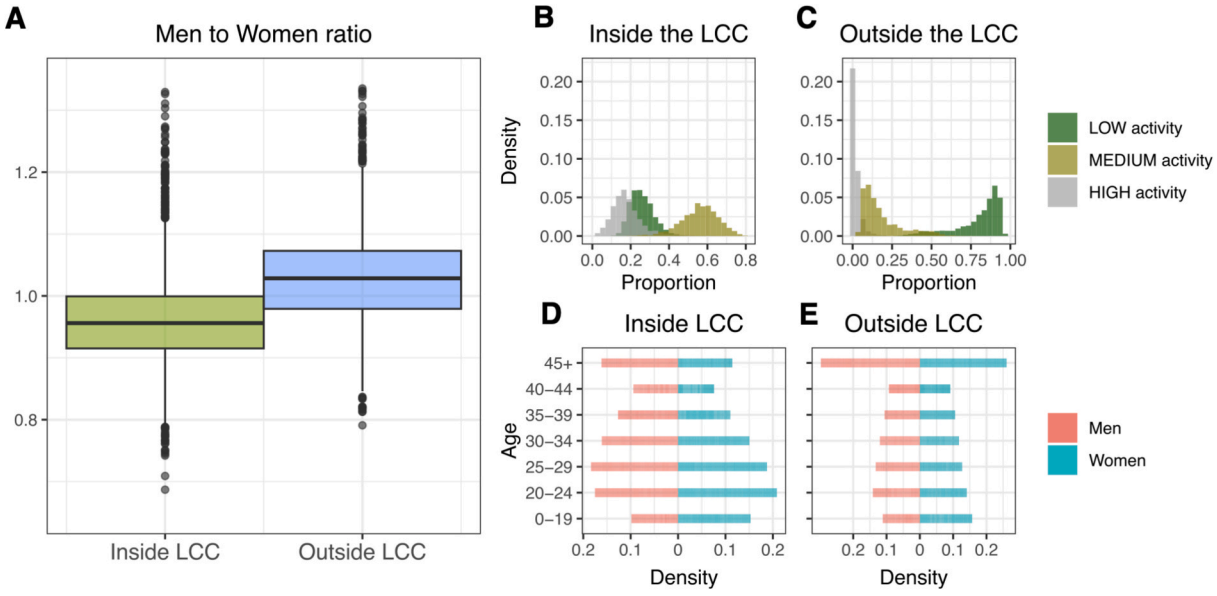

Fig. 3. Demographic analysis of the large connected component versus the rest of the partnership network across different calibrated runs: A Ratio of men to women, B – C, distribution of risk activity within and outside the large connected component, and D – E age pyramid for men and women within and outside the large connected component. Result for all the communities aggregated.

Table 1
Risk factors for being in the lcc, multivariable and univariable regressions. For multivariable regressions, the reference is women, 30 to 40*years*, with medium sexual activity.

	Odds ratio for each risk factor for being in the large connected component	
	Multivariable analysis	
Men	0.886	[0.884–0.887]
< 20	0.413	[0.412–0.414]
20–30	1.050	[1.049–1.051]
40+	0.396	[0.396–0.397]
high sexual activity	3.004	[2.997—3.010]
low sexual activity	0.052	[0.052–0.053]
	Univariable analysis	
Men	0.938	[0.937,0.939]
15 – 20	0.881	[0.879,0.882]
20–30	1.153	[1.152,1.154]
40+	0.516	[0.515, 0.516]
high sexual activity	3.122	[3.115,3.129]
low sexual activity	0.059	[0.059,0.060]

missions in serodiscordant couples. Fig. 4 shows the incidence and the prevalence across the overall network versus the lcc. Supplementary Figs. 3 and 4 depict the same analysis disaggregated by country. Across all communities, the percentage of incident cases emerging indicates that being in the lcc is amongst the most important risk factors in determining one's chances of acquiring HIV. To show this, we performed logistic regression on determinants of acquiring HIV, looking specifically at the effects of sexual activity level, age, and being in the component at the time of HIV acquisition (Table 2). In both univariable and multivariable models, belonging to the large connected component is associated with the highest risk of infection (odds ratio versus not being in the large connected component 7.43). Being part of the most sexually active of the population is also associated with a 2.29-fold increase in risk of infection compared to the population in the medium level, indicating that high sexual activity is also an important risk factor for infection.

Looking in more detail at the offspring distribution, that is, the distribution of the number of onward transmissions per index case, we observe homogeneous distributions (Fig. 5). On average, each transmission results in 0.77 onward transmissions suggesting that incidence is declining. If we focus only on transmissions which lead to at least one onward transmission over a lifetime, the offspring distribution has a mean of 1.77, with a variance of 1.31 and a maximum of $m = 11$ onward transmissions, indicating that large numbers of transmissions from a single source are not observed.

4. Discussion

In this work, we investigated sexual networks emerging from the output of PopART-IBM, an agent-based model calibrated to data from multiple sources on HIV prevalence, awareness of HIV status, ART status, and viral suppression for HPTN 071 (PopART) study communities. The model explicitly defines the rules of sexual partnership formation and dissolution, implying the emergence of a dynamic sexual network. HIV transmission is modelled as a process spreading through heterosexual serodiscordant couples, meaning that the transmission network is embedded in the partnership network. Hence, analysing the partnership network derived from this model is highly informative in understanding epidemic patterns across regions in sub-Saharan Africa.

By looking at the outputs of the calibrated model across all the communities of the PopART study in Zambia and South Africa, we found that all sexual partnership configurations compatible with a generalised epidemic in such regions contain large connected components. A significant fraction of the sexually active population belongs to a single large connected component of the network, with half of the simulations outputting networks with a large connected component embracing more than 40% of the whole network. It is important to note that this effect is not due solely to the structure of the model, as it is flexible enough to generate a wide variety of networks, many of which are missing such large connected component. Instead, it is only when the model is calibrated to data that this large connected component emerges.

A demographic and behavioural analysis of these networks reveals that the population within the large connected component has specific characteristics: age between 20 and 40, and more sexually active than

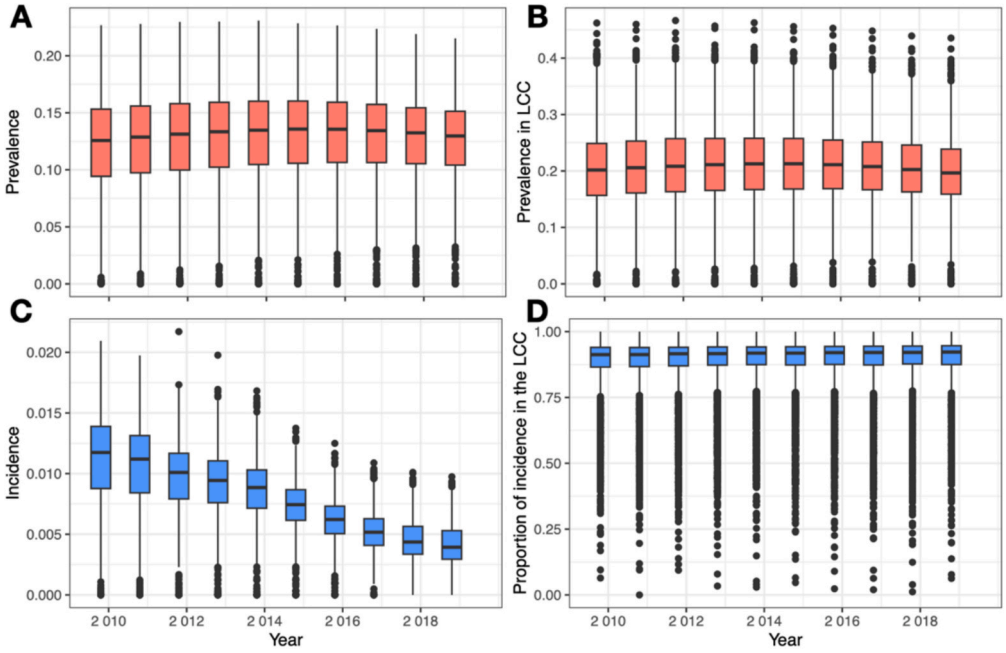

Fig. 4. A Prevalence in the whole network and B in lcc. C Overall incidence per person year, and D, proportion of annual incidence occurring in the large connected component. Results obtained aggregating all the communities after performing 100 simulations on each.

Table 2
Risk factors for HIV acquisition, multivariable and univariable regressions. For multivariable regressions, the reference is women, 30 to 40*years*, with medium sexual activity. The last line refers to univariable regression with being in the large connected component as the only determinant to transmission.

	Relative Risk of HIV acquisition	
	Multivariable analysis	
Males	0.670	[0.667, 0.674]
20–30	1.256	[1.248, 1.266]
40+	0.856	[0.848, 0.864]
high risk	2.118	[2.104, 2.130]
low risk	0.539	[0.533, 0.543]
Being in the large connected component	4.126	[4.085, 4.168]
	Univariable analysis	
Males	0.727	[0.723, 0.731]
<20	1.116	[1.107, 1.126]
20–30	1.434	[1.108, 1.126]
40+	0.628	[0.622, 0.634]
high risk	2.296	[2.282, 2.310]
low risk	0.275	[0.273, 0.277]
Being in the large connected component	7.431	[7.366, 7.500]

the average person. Although the network is dynamic, sexually active individuals typically remain in the large connected component for many years, implying a higher risk of acquiring HIV, as the analysis of the transmission network reveals that around 90 % of the incident cases over a 10-year period happens among people in the large connected component. The importance of patterns of sexual behaviour in driving the epidemic of HIV have long been investigated in the literature (Donovan and Ross 2000; Moody, Adams, and Morris 2017; Anderson et al. 1991).

The study has several limitations. First, the network is not directly observed in or reconstructed from the available data, but it is an emergent property of the combination of data analysis and assumptions of the model. However, it is encouraging that all the calibrated runs across all the communities exhibit the same pattern across a calibration with a wide range of parameters. Second, in communities where incidence is higher, we observe relatively larger connected components and higher partnership formation rates, suggesting that the model may not be fully identifiable as it tends to explain higher incidence with higher sexual activity in the community. Third, the model is calibrated to population-cohort data, which only include people aged 18–44. Fourth, the model accounts only for sexual transmission in heterosexual individuals, which is a limitation as including other routes of transmission may have an impact also on the heterosexual relationship network. Fifth, the model does not consider some sub-populations, such as sex workers or MSM, who may less frequently engage with research. These subpopulations were not sampled sufficiently during PC rounds or individuals belonging to these groups did not disclose themselves as such. Unsampled individuals may have more sexual partners than sampled individuals and thus become “hubs” in the sexual network, therefore contributing more to community transmission. Further, marriage, religion, and other socioeconomic factors are not modelled, although these may contribute to shaping the dynamics, possibly mediated by sexual

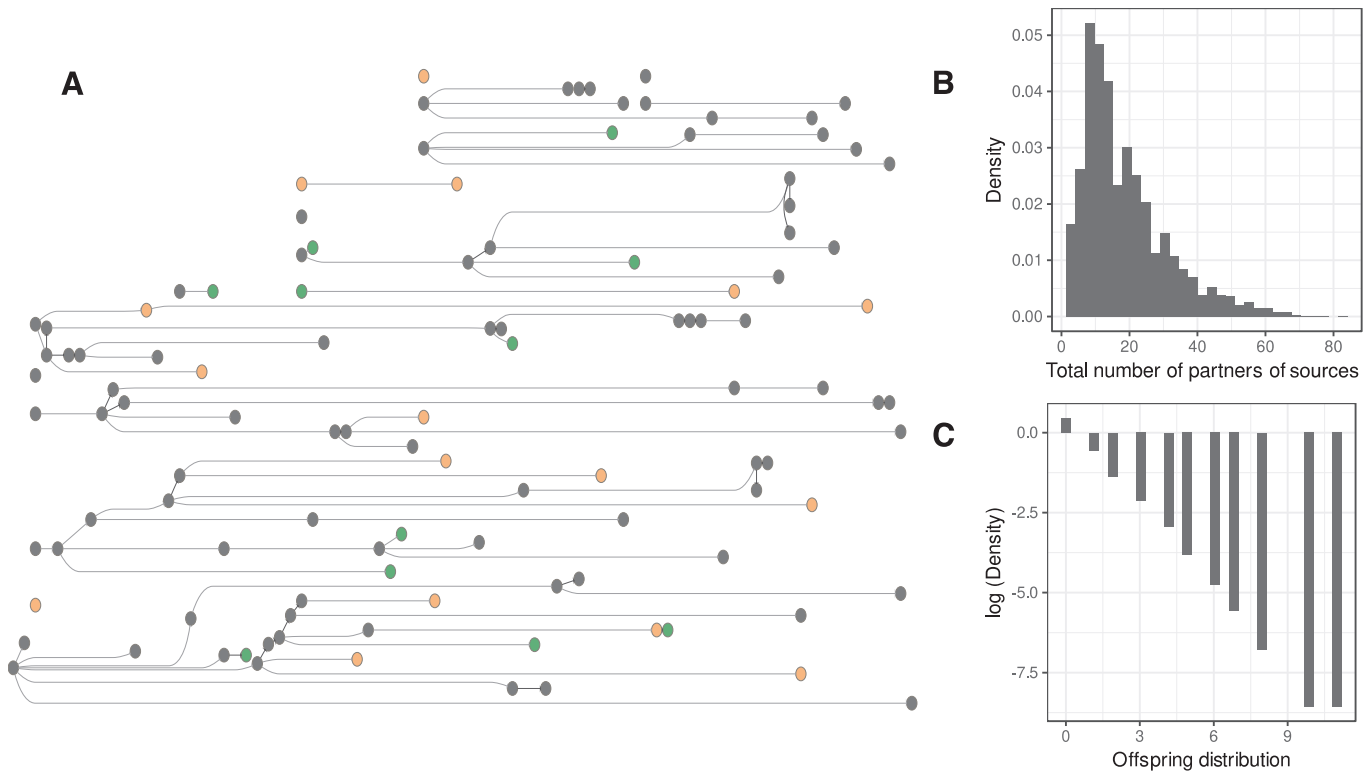

Fig. 5. A Selection of branches in the transmission networks sampled over ten years for a randomly chosen calibrated run, nodes colored by activity level. Time flows from left to right. Each node is an individual, and each link is a transmission event. The color of each node indicates the sexual activity class (green – low, gray – medium, yellow – high). B distribution of the total lifetime partners for individuals with at least one onward transmission in the same run. C logarithm of the offspring distribution of HIV sources.

behaviour or circumcision status. Another limitation is that the data on which the parameters of the networks are fitted may be biased. Under-reporting among people who report few partners due to social desirability bias and the relative undersampling of highly sexually active people (e.g. female sex workers and their clients) may impact relevant quantities such as the size of the lcc. This can happen due to the model needing a lcc to accommodate an HIV epidemic and therefore compensating the absence of hubs with a higher average number of partners. However, after looking at the distribution of partners in a heterosexual population (Schneeberger et al., 2004), around 4 people in 10,000 are expected to have over 50 relationships per year. Further, most of these would be one-off, not contributing much to the over-all structure of the network.

Finally, different models for partnership formation/dissolution were not explored, e.g. there are no explicitly modelled commercial sex worker, nor one-off relationships, which may have an impact both on the HIV dynamics and on the large connected component. In future studies, we aim at improving the model by addressing these limitations.

In summary, using agent-based models to investigate properties that are not directly observed in the data may be a powerful tool to understand the determinants of risk of HIV acquisition and generate insights that can be useful for policy makers. The main result of this work is that the greatest risk of HIV acquisition is related to being in a part of the sexual network that is densely connected. While the odds of being in a large connected component depend on age and sexual behaviour (young and sexually active individuals), there is a significant part of the population that does not have any of these risk factors, but whose partners happen to be in the large connected component. This is a possible explanation of why the epidemic in some countries in southern and eastern Africa is generalised. One of the most pressing questions for HIV prevention is whether epidemic patterns observed in Sub-Saharan Africa are driven by core groups (Lowndes et al. 2002; Boily et al., 2002). A core group is usually thought of as a relatively small subpopulation, such as commercial sex workers (Lowndes et al. 2002), that accounts for most of the new infections in a population. With PopART-IBM, we showed that according to available data, a large part of the population is involved in the epidemic dynamics. While this does not mean that there are no sub-populations whose risk of acquiring HIV is much greater than the general population, and whose contribution to the overall incidence is disproportionate, a large part of the population at risk of HIV acquisition consists of young and highly sexually active individuals. Taken together, these findings suggest that populations consisting of young and highly sexually active individuals should be given high priority when designing or deploying interventions.

CRediT authorship contribution statement

Francesco Di Lauro: Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **William J.M. Probert:** Writing – review & editing, Data curation. **Michael Pickles:** Writing – review & editing, Data curation. **Anne Cori:** Writing – review & editing, Software, Formal analysis, Data curation. **Robert Hinch:** Writing – review & editing, Software. **Luca Ferretti:** Writing – review & editing, Validation. **Jasmina Panovska-Griffiths:** Writing – review & editing. **Lucie Abeler-Dörner:** Writing – review & editing, Supervision. **Rory Dunbar:** Writing – review & editing, Data curation. **Peter Bock:** Writing – review & editing, Data curation. **Deborah J. Donnell:** Writing – review & editing. **Helen Ayles:** Writing – review & editing, Data curation. **Sarah Fidler:** Writing – review & editing. **Richard Hayes:** Writing – review & editing. **Christophe Fraser:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

PANGEA is funded by the Bill & Melinda Gates Foundation (consecutive grants OPP1084362 and OPP1175094).

MP is supported partly by the HPTN Modelling Centre, which is funded by the U.S. National Institutes of Health (NIH UM1 AI068617) through HPTN and by the UNAIDS. MP acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1), funded by the UK Medical Research Council (MRC).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jtbi.2025.112218.

References

Anderson, R.M., May, R.M., Boily, M.C., Garnett, G.P., Rowley, J.T., May, R.M., 1991. The spread of HIV-1 in Africa: sexual contact patterns and the predicted demographic impact of AIDS. Nature 352 (6336), 581–589.

Bellan, J., Galvani, S.E., Dushoff, J., 2015. Reassessment of HIV-1 acute phase infectivity: accounting for heterogeneity and study design with simulated cohorts. PLoS Med. 12 (3), 1–28.

Boily, M.C., Lowndes, C., Alary, M., 2002. The impact of HIV epidemic phases on the effectiveness of core group interventions: insights from mathematical models. Sex. Transm. Infect. 78 (suppl 1), 1–i90.

Di Lauro, F., Croix, J.C., Dashti, M., Berthouze, L., Kiss, I.Z., 2020. Network inference from population-level observation of epidemics. Sci. Rep. 10 (1), 18779.

Donovan, B., Ross, M.W., 2000. Preventing HIV: determinants of sexual behaviour. Lancet 355 (9218), 1897–1901.

Hallett, T.B., Gregson, S., Lewis, J.J.C., Lopman, B.A., Garnett, G.P., 2007. Behaviour change in generalised HIV epidemics: impact of reducing cross-generational sex and delaying age at sexual debut. Sex. Transm. Infect. 83 (suppl 1), 1–i54.

Hayes, R.J., Donnell, D., Floyd, S., Mandla, N., Bwalya, J., Sabapathy, K., Yang, B., et al., 2019. Effect of universal testing and treatment on HIV incidence — HPTN 071 (PopART). N. Engl. J. Med. 381 (3), 207–218.

Hayes, R., Ayles, H., Beyers, N., Sabapathy, K., Floyd, S., Shanaube, K., Bock, P., et al., 2014. HPTN 071 (PopART): rationale and design of a cluster-randomised trial of the population impact of an HIV combination prevention intervention including universal testing and treatment–a study protocol for a cluster randomised trial. Trials 15 (1), 1–17.

Helleringer, S., Kohler, H.-P., 2007. Sexual network structure and the spread of HIV in Africa: evidence from Likoma Island. Malawi. AIDS. 21 (17), 2323–2332.

Hudson, C.P., 1996. AIDS in rural Africa: a paradigm for HIV-1 prevention. Int. J. STD AIDS 7 (4), 236–243.

Justman, J.E., Mugurungi, O., El-Sadr, W.M., 2018. HIV population surveys—bringing precision to the global response. N. Engl. J. Med. 378 (20), 1859–1861.

Keeling, M., 2005. The implications of network structure for epidemic dynamics. Theor. Popul Biol. 67 (1), 1–8.

Kiss, I.Z., Green, D.M., Kao, R.R., 2006. The network of sheep movements within Great Britain: network properties and their implications for infectious disease spread. J. R. Soc. Interface 3 (10), 669–677.

Lenormand, M., Jabot, F., Deffuant, G., 2013. Adaptive approximate Bayesian computation for complex models. Comput. Stat. 28 (6), 2777–2796.

Lowndes, C.M., Alary, M., Meda, H., Gnintoungbé, C.A.B., Mukenge-Tshibaka, L., Adjovi, C., Buvé, A., et al., 2002. Role of core and bridging groups in the transmission dynamics of HIV and STIs in Cotonou, Benin, West Africa. Sex. Transm. Infect. 78 (suppl 1), 1–i77.

Magosi, L.E., Zhang, Y., Golubchik, T., DeGruttola, V., Tchetgen Tchetgen, E., Novitsky, V., Moore, J., et al., 2022. Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial - BCPP/Ya Tsie Trial. Elife 11.

Moody, J., Adams, J., Morris, M., 2017. Epidemic potential by sexual activity distributions. Network Science. 5 (4), 461–475.

Newman M. Networks. Oxford University Press. 2018.

Perriat, D., Balzer, L., Hayes, R., Lockman, S., Walsh, F., Ayles, H., Floyd, S., et al., 2018. Comparative assessment of five trials of universal HIV testing and treatment in Sub-Saharan Africa. J. Int. AIDS Soc. 21 (1).

Pickles, M., Cori, A., Probert, W.J.M., Sauter, R., Hinch, R., Fidler, S., Ayles, H., et al., 2021. PopART-IBM, a highly efficient stochastic individual-based simulation model of generalised HIV epidemics developed in the context of the HPTN 071 (PopART) trial. PLOS Comput. Biol. 17 (9).

Probert, W.J.M., Sauter, R., Pickles, M., Cori, A., Bell-Mandla, N.F., Bwalya, J., Abeler-Dörner, L., et al., 2022. Projected outcomes of universal testing and treatment in a generalised HIV epidemic in Zambia and South Africa (the HPTN 071 [PopART] trial): a modelling study. Lancet HIV 9 (11), e780.

Read, J.M., Keeling, M.J., 2003. Disease evolution on networks: the role of contact structure. Proc. R. Soc. Lond. B 270 (1516), 699–708.
Robinson, S.E., Everett, M.G., Christley, R.M., 2007. Recent network evolution increases the potential for large epidemics in the British cattle population. J. R. Soc. Interface 4 (15), 669–674.
Schaub, M.T., Segarra, S., Tsitsiklis, J.N., 2020. Blind identification of stochastic block models from dynamical observations. SIAM J. Math. Data Sci. 2 (2), 335–367.
Tanser, F., Bärnighausen, T., Hund, L., Garnett, G.P., McGrath, N., Newell, M.-L., 2011. Effect of concurrent sexual partnerships on rate of new HIV infections in a high-prevalence, rural South African population: a cohort study. Lancet 378 (9787), 247–255.
Tanser, F., Hosegood, V., Bärnighausen, T., Herbst, K., Nyirenda, M., Muhwava, W., Newell, C., Viljoen, J., Mutevedzi, T., Newell, M.-L., 2008. Cohort profile: Africa centre demographic information system (ACDIS) and population-based HIV survey. Int. J. Epidemiol. 37 (5), 956–962.
Corsi, D.J., Neuman, M., Finlay, J.E., Subramanian, S.V., 2012. Demographic and health surveys: a profile. Int. J. Epidemiol. 41 (6), 1602–1613. https://doi.org/10.1093/ije/dys184. Epub 2012 Nov 12 PMID: 23148108.
Vijver DAMC vd, Prosperi MCF, Ramasco JJ. Transmission of HIV in sexual networks in Sub-Saharan Africa and Europe. Eur. Phys. J. Spec. Top. 2013;222(6):1403–1411.
Schneeberger, A., Mercer, C.H., Gregson, S.A., Ferguson, N.M., Nyamukapa, C.A., Anderson, R.M., Johnson, A.M., Garnett, G.P., 2004. Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in Britain and Zimbabwe. Sex. Transm. Dis. 31 (6), 380–387. https://doi.org/10.1097/00007435-200406000-00012. PMID: 15167650.
Watts, C.H., May, R.M., 1992. The influence of concurrent partnerships on the dynamics of HIV/AIDS. Math. Biosci. 108 (1), 89–104.