RESEARCH ARTICLE

Network centrality for the identification of biomarkers in respondent-driven sampling datasets

Jacob Grubb, Derek Lopez, Bhuvaneshwar Mohan, John Matta◉*

Computer Science Department, Southern Illinois University Edwardsville, Edwardsville, IL, United States of America

* jmatta@siue.edu

Abstract

Networks science techniques are frequently used to provide meaningful insights into the populations underlying medical and social data. This paper examines SATHCAP, a dataset related to HIV and drug use in three US cities. In particular, we use network measures such as betweenness centrality, closeness centrality, and eigenvector centrality to find central, important nodes in a network derived from SATHCAP data. We evaluate the attributes of these important nodes and create an exceptionality score based on the number of nodes that share a particular attribute. This score, along with the underlying network itself, is used to reveal insight into the attributes of groups that can be effectively targeted to slow the spread of disease. Our research confirms a known connection between homelessness and HIV, as well as drug abuse and HIV, and shows support for the theory that individuals without easy access to transportation are more likely to be central to the spread of HIV in urban, high risk populations.

Introduction

In this paper we utilize several types of network centrality measures, including betweenness, closeness, and eigenvalue, to find nodes central to the structure of the SATHCAP Referral Network, a network based on concurrent drug use and sexual activity. Epidemiological theory suggests that interventions targeting these central nodes are more effective in stopping the spread of a disease through a network than interventions involving other nodes [1, 2]. By finding and examining these central nodes, we hope to find a set of shared attributes, or biomarkers. Knowledge of these biomarkers can be used to design targeted intervention campaigns to make diseases such as HIV less likely to spread through high risk environments.

The importance of learning about HIV transmission via tools such as network science is great. HIV began as an epidemic in 1981 and continues to impact many segments of society, particularly populations considered *hidden* or *hard-to-reach*, such as minorities and men who have sex with men (MSM). The disease remains stigmatized, such that 1 in 7 people who have HIV do not know it [3], and half of minority MSM will become HIV-infected in their lifetime [4]. Minority women suffer worse health outcomes with HIV than other women [5].

Surveys conducted using respondent-driven sampling (RDS) attempt to discover information specifically about a hidden population through the use of peer recruitment to find participants [6]. The SATHCAP RDS [7] focuses on low-income, minority MSM and injected drug users. A large amount of data was collected from each participant, making this study a particularly rich source of information about a hidden population that is likely to be affected by HIV. The SATHCAP study was conducted between 2006 and 2008 in the US cities of Chicago, Los Angeles, and Raleigh, NC. The 4,688 participants were asked almost 1500 questions concerning their sexual and drug-related habits, as well as demographic and other information. The nature of the RDS sampling in this study is that new recruits had participated in potential HIV-spreading behavior with their recruiter. The derived network is therefore representative of one through which HIV and other diseases could have spread. Our methodology finds the three largest connected components within each city and looks for shared attributes between the central-most nodes within these components.

If properties of actors susceptible to HIV were analyzed through a statistically random survey of the general population, it is likely that few members of the hidden populations of interest would be found, because of overall low prevalences. Use of an RDS survey has two benefits. First, sampling only the hidden population allows an analysis within that population, giving more detailed results when determining attributes of high risk individuals. Second, an RDS survey preserves the connections between individuals, which is another source of information that is lost in a large independent survey.

In an RDS survey the initial samples (called *seeds*) are not random, but if there are enough recruitment waves the surveyed population becomes statistically independent of the seeds [6]. An RDS survey relies on *homophily* in recruitment, which is the property that people in the studied hidden populations know each other (while remaining mostly unknown to researchers). Because of homophily, the construction of the SATHCAP network is not necessarily random [8], but instead has a higher probability of recruitment of the target population. The RDS network obtained is a series of trees, which is a by-product of the recruitment process. While the underlying network may have loops, studies have shown that sexual networks have few loops [9], including the HIV transmission network of a large US city [10]. Because the network underlying SATHCAP is likely scale free [11] (as are many sexual networks, for example [12]), high degree nodes (or hubs) are important. The probability that the RDS sample contains hubs is great—much greater than the probability that they would be included in a general-population random sample.

Finding and helping vulnerable hidden populations has previously shown great benefits. Due to extensive preventative efforts, HIV diagnosis rates from 2010 to 2015 decreased in the categories of MSM and persons who inject drugs. However, diagnosis rates still increased in the more specific (and hidden or hard-to-reach) categories of black individuals and MSM aged 25-34 [13]. Recent studies have shown that viral suppression (achieved through highly active antiretroviral therapy, or HAART) prevents sexual transmission of HIV [14]. Once at-risk groups are identified, such as through network science techniques, preventative interventions can be attempted such as HAART for infected individuals, PrEP for uninfected individuals [15], and educational interventions for all.

This paper is an extension of [11]. The previous paper used only betweenness centrality as a metric for importance. This paper improves upon the previous paper by also examining closeness centrality and eigenvector centrality. Each method of centrality discovers a different set of central nodes, each of which has their own set of notable attributes. We further improve upon the previous work by distributing the set of central attributes between the three largest components, taking the top five central nodes from each component to yield a total of 15 nodes, rather than taking the top 10 most central nodes from each city. Because of the homophily

involved in the data collection, nearby nodes within the same component are more likely to share similar attributes. Choosing nodes from different components prevents the largest components from overwhelming the other large components and gives a more diverse sample of central nodes across the city networks.

The data have been obtained through the National Addiction and HIV Data Archive Program (NAHDAP), accessible online https://www.icpsr.umich.edu/icpsrweb/NAHDAP/index.jsp. This research was conducted under the approval of the Southern Illinois University Edwardsville IRB.

Related work

The SATHCAP dataset was the basis for several papers published in a special issue of the Journal of Urban Health in 2009 [7]. One paper from that special issue that focuses on network interpretation is [16]. In that paper Youm et al. identify sets of "bridging individuals" between communities in Chicago. That paper also identifies hidden communities that have an impact on the spread of the HIV despite low disease incidence rates. Because of their low overall incidence rates, these communities would not have been discovered through a general population survey. However, targeted interventions within these specific communities would materially reduce the transmission rate through the entire city of Chicago. In other work, Ober et al. [17] use the SATHCAP dataset to identify factors associated with stimulant drug use during sexual activity between older, low income males.

Network analysis of spreading has been conducted in many contexts, such as in terms of social contacts [18] and rodent infestations [19], among others. Because contacts are generally known and can be traced, network analysis is particularly suited to sexually transmitted diseases, such as gonorrhoea [20] and chlamydia [21] in addition to HIV [22]. Liljeros et al. [23] provide an introduction into the use of network theory to describe sexual interactions and the application of this theory to the spread of venereal diseases. An early example of the use of network theory to represent the spread of HIV can be found in [24], which provides a series of mathematical formulas to predict the spread of disease in heterogeneous sexual networks.

Graph-based aspects of respondent driven sampling, and the fact that "data collected in ordinary RDS studies contain information about the structure of the respondents' social network" are discussed in [25]. The three network centrality measures considered in this paper are well-known and have previously been used many times in the context of analyzing disease spread. Betweenness centrality is often used to find topologically important nodes relating to network resilience and spreading [26], including the spread of diseases like HIV [27]. Closeness centrality has been found to be influential in models of disease spreading in pig populations [28]. In [29] it is found that SARS-CoV-2 spreads to Brazilian cities more quickly based on the closeness centrality of the city's airport. Eigenvector centrality is used in many instances to study the spread of disease in epidemics [30, 31], and in [32] it is shown that "eigenvector centrality approximately quantifies the risk of a node to become infected" with COVID-19. In [33] it is shown that the eigenvector centrality values of the seed nodes in an Italian cattle network are strongly correlated with the extent of the spread of an epidemic through the network. In a bioinformatics example of its use, eigenvector centrality has connections to brain network alterations in Alzheimer's disease [34].

Respondent driven sampling (RDS) was first presented as part of an AIDS prevention initiative, as a method for interviewing subjects independent of the original sample subject [6]. By utilizing chains of peer recruitment combined with Markov modeling, RDS is able to access remote populations while theoretically reducing sampling bias to a reliably low level [35].

Methodology

Respondent driven sampling

RDS is a data collection technique in which samples are generated from a random walk along nodes in the underlying network. Sampling probability is proportional to the node's degree [36]. There are several assumptions necessary to ensure the independence of samples in the RDS process. The members of populations sampled must be able to identify each other, and preferably will have acquaintanceships that form a connected network. The size of data collected must be small relative to the overall size of the population such that the pool of recruits is not quickly exhausted. Participants must be able to choose new recruits randomly, and participants must be able to accurately report the number of acquaintanceships they have. The weakness of RDS lies in the degree to which these assumptions may be false [8]. The strength of RDS is in obtaining information about hidden populations that may be impossible to obtain through standard random surveys.

It is shown in [11] that the underlying network of SATHCAP is scale free. This implies that the network has hubs [37], and the distance from a random node to a hub in a scale free network is short (in theory, of the order of the log of the size of the network [38]). RDS (and snowball-style sampling methods in general) tends to choose hubs [39, 40]. This fact is used in the analysis of RDS data. For example, Successive Sampling Population Size Estimation (SS-PSE) is a technique developed by Gile and Handcock [36] for estimating the size of the underlying network based on an RDS sample. This technique assumes that high degree nodes are sampled early in the process, and if high degree nodes remain after several waves of recruitment, the size of the hidden population must be correspondingly larger. Also, because of its propensity to select hubs, RDS has been suggested as a way to find high degree nodes for immunization [41].

Data parsing and cleaning

The original SATHCAP dataset contained a total of 4688 participants who were asked 1493 questions. The answers given are referred to here as *features* or *attributes* of the participants. Peer recruitment into the study happened via coupons which were given to seed participants to distribute to possible study recruits. The coupons were color-coded, based on the relationship between the recruiter and recruitee, with different colors representing sexual connections and drug-using connections. Information from the coupon numbers, city codes, and colors were used to create unique identifiers for each participant. This enabled each participant to be traced to a position in the recruitment network.

Due to the length of the survey and the nature of the questions, respondents were given the option to decline to answer questions. This meant that there was a great deal of missing data. Over 40% of features (n = 1352/1488) are missing more than 40% of observations. Features missing more than a given threshold percentage of observations were removed, where the optimal threshold has a high number of features with a low percentage of missing observations. Features having at least 94% of observations were retained, except features consisting of metadata. This left a total of 80 features.

One-hot encoding was used to convert attributes without an ordinal relationship into separate binary attributes. For example, a question about income with possible answers of *high*, *medium*, and *low* is converted into three questions with yes/no answers: *Is income high?, Is income medium?, Is income low?*. Thus the answers to the questions become features of the participants, such as *low income*. The end result after one-hot encoding of multi-valued attributes was a set of 141 features.

City network creation

The city networks were created using the participant identifier numbers. Each participant is a node in the network, and links were created between each node and the nodes it recruited. The result is a forest of tree graphs. Many of the seed nodes did not recruit additional participants. In fact, 255 out of 412 connected components in the final network are of size 1 or 2, indicating a situation in which a seed node recruited no one or only 1 participant. Successful chains of recruitment have between 30 and 949 nodes. For our analysis, we look at the three largest components within Los Angeles, Chicago, and Raleigh-Durham, for a total of 9 distinct graphs across three cities. Due to the structure of recruitment, there are no cycles in the network (a participant could not recruit someone who had already been recruited), and the maximum degree of a node is 7 (a participant could recruit a maximum of 6 others).

Calculation of centrality

As an improvement upon [11], we include two new forms of centrality in addition to the previous betweenness centrality. This work includes the use of closeness centrality and eigenvector centrality as metrics of importance. Each of these centrality measures is concerned with different network properties. Calculated example values for a simple tree network are shown in Fig 1.

The eigenvector centrality of a node is based on the node's degree, and is a numerical representation of the "popularity" of a node, where nodes that are connected with other "popular" or high degree nodes are given higher centrality scores. As such, it is a useful measure for determining highly influential (and influenced) nodes in social networks [42, 43]. This measure is calculated by finding the principal eigenvector of the adjacency matrix of the graph. By finding "popular" nodes within SATHCAP, we highlight sets of nodes with a high probability of infection based on their proximity to infected nodes. Eigenvector centrality does this without relying exclusively on degree centrality, which is problematic on a graph whose degree is artificially limited, such as SATHCAP. In the Fig 1 example, nodes 7 and 3 have the highest eigenvector centrality, and are also the highest degree nodes. Despite having a low degree, node 6 has a high eigenvector centrality because of its proximity to nodes 7 and 3. Nodes 8, 9 and 10 have a medium eigenvector centrality, despite having a low degree of 1, based only their proximity to high-degree node 7. Compare this to node 0, also with degree 1, which has a
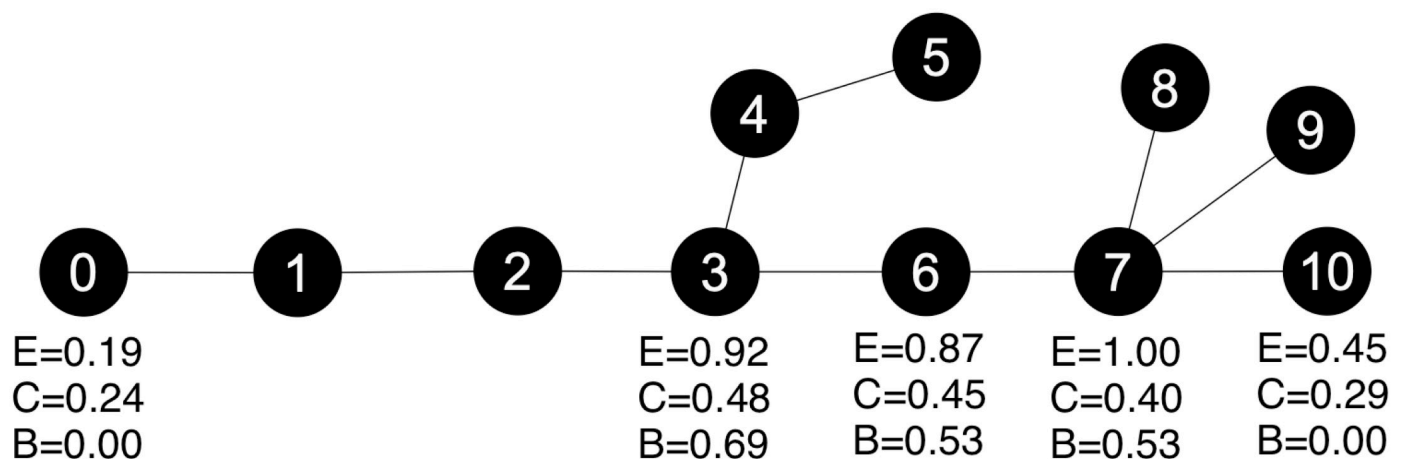
Fig 1. Values for different centrality measures are shown for nodes in a tree. *E* represents eigenvector centrality, *C* is closeness centrality, and *B* is betweenness centrality.

https://doi.org/10.1371/journal.pone.0256601.g001

low eigenvector centrality. It can be seen that eigenvector centrality is a meaningful measure, even in a small tree example.

The closeness centrality [44] of a node is defined as the inverse of the sum of distance of the shortest paths from that node to every other node in the connected component. Mathematically, normalized closeness centrality is expressed as

$$c(v) = \frac{N}{\sum_y d(y, v)} \tag{1}$$

where N is the number of nodes, and $d(y, v)$ represents the distance from node y to node v. Closeness centrality is useful for identifying nodes that are near to other nodes. Within SATH-CAP, these nodes would represent a hazard if infected, as the short distance to every other node would potentially represent an easier path for transmission. In Fig 1 nodes 3 and 6 have the highest closeness centrality. Node 7 has a comparatively lower closeness centrality, despite its high degree. Node 0 is furthest from other nodes and therefore has the lowest closeness centrality.

The betweenness centrality [45, 46] of a node is based on the shortest paths property of a graph, and is defined as the proportion of paths that traverse that node when calculating the shortest paths from all nodes to all other nodes. A mathematical formula describing this measure can be seen in Eq 2, where σ_{st} represents the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ represents the number of those paths that contain node v.

$$b(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2}$$

Betweenness centrality is a valuable metric within the SATHCAP respondent driven sampling network, as it highlights the nodes that lie on the shortest and most direct paths for the transmission of disease. Node 3 has the highest betweenness centrality in Fig 1. As an example of the differences of the three centrality measures, note that while nodes 8, 9, and 10 have relatively high eigenvector centrality, they do not lie on any transmission paths, and therefore have a betweenness centrality of zero.

Identification of central nodes

After calculating the betweenness, closeness, and eigenvector centrality scores for each node within the SATHCAP referral network, we need to identify the nodes that are most central.

In [11], we used the trivial solution of taking the ten nodes with the highest score in each city. While this is logical, difficulties arise when plotting these nodes on the underlying graph. Due to large discrepancies in the size of components, many to all of the central nodes would appear in a single component, frequently connected together in a chain. This methodology resulted in discarding other large, important components that were overshadowed by the size of the largest component.

To combat this overshadowing, we instead take our central nodes from a variety of components. We find the five nodes with the highest centrality scores from each of the three largest components in each city. This yields a total of 15 central nodes from each city, better distributed throughout each of the large components.

Calculation of attribute exceptionality

Once the set of central nodes, c, within each city has been identified, we examine the attributes, i, of the respondent associated with each central node. We compare the value of each attribute

to the average value of that attribute for the city it belongs to. The 'city average' for each variable is taken to be the average value of all respondents from that city within the SATHCAP dataset. We note if an attribute appears more than two standard deviations away than the city average, indicating, based on a 95% confidence interval, that the value is an outlier from the average response. If an attribute appears deviant in a large number (at least 4 of 15) of central nodes, we consider that attribute to be "exceptional," and give it an exceptionality score equal to the number of central nodes that possessed that deviant attribute, c_i, divided by the total number of central nodes, such that $exceptionality = \frac{|c_i|}{|c|}$. We repeat this calculation for each city and centrality type to find a set of attributes for each city-centrality combination.

Results

Figs 2–4 highlight the nodes identified by each centrality method within a single component of each city. It is apparent that there are differences in the set of nodes identified by each centrality method. High betweenness nodes are highlighted in red, nodes with high closeness centrality are green, and nodes with high eigenvector centrality are purple. The seed node for each component is yellow. Each subfigure is labeled by its corresponding centrality method. By looking at the shared sets of attributes within these nodes and counting the number of times those attributes fall more than two standard deviations from the city average, we find a set of shared, exceptional attributes that best describe the central nodes.
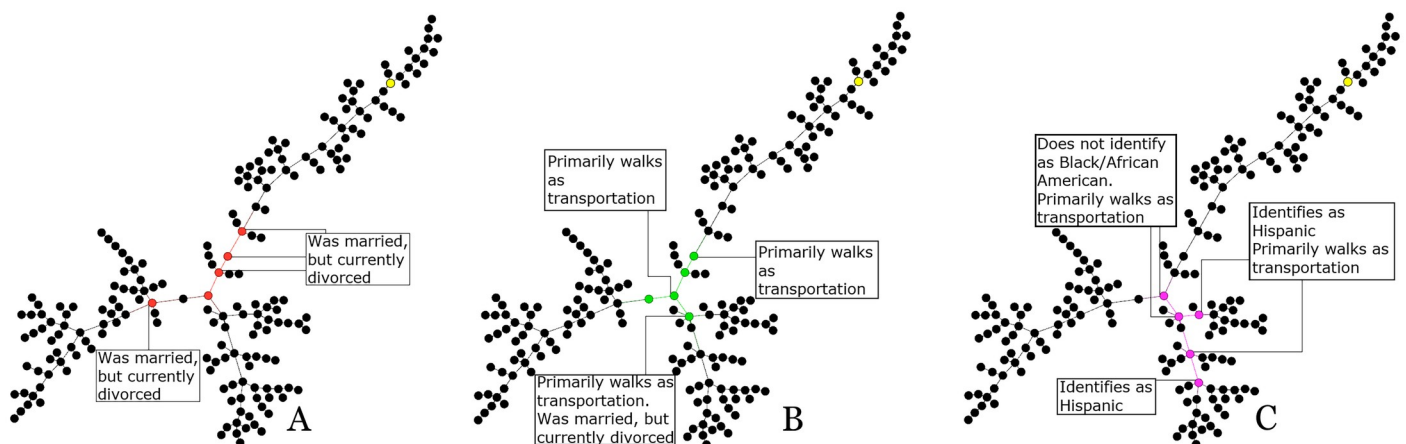
Fig 2. One component from the Chicago referral network, with nodes highlighted by centrality. **A**. Betweenness. **B**. Closeness. **C**. Eigenvector.

https://doi.org/10.1371/journal.pone.0256601.g002
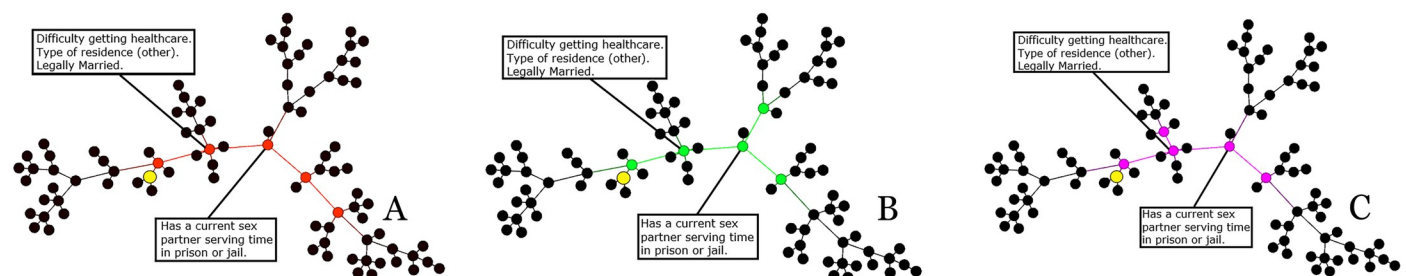
Fig 3. One component from the Los Angeles referral network, with nodes highlighted by centrality. **A**. Betweenness. **B**. Closeness. **C**. Eigenvector.

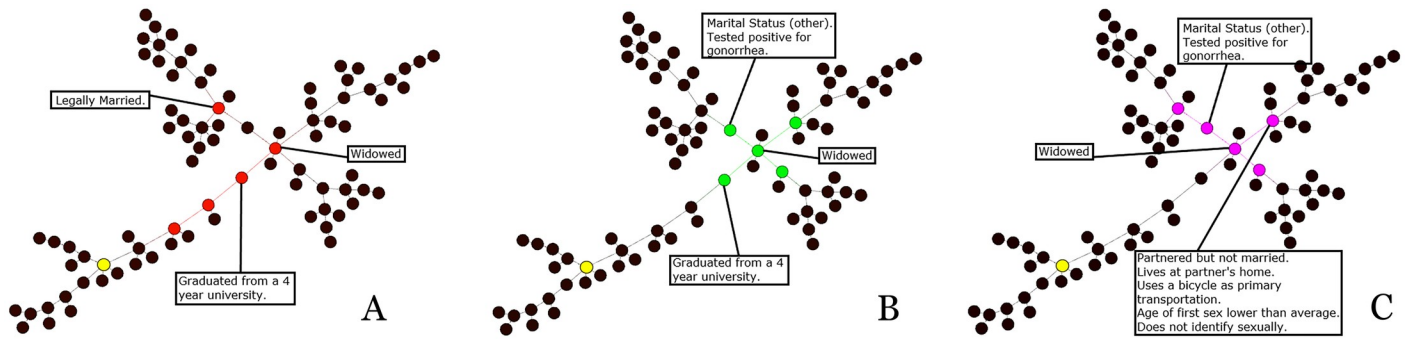https://doi.org/10.1371/journal.pone.0256601.g003

Fig 4. One component from the Raleigh-Durham referral network, with nodes highlighted by centrality. A. Betweenness. **B**. Closeness. **C**. Eigenvector.

The three centrality methods were run on the three largest connected components in the Chicago network. Results for one of the Chicago components are shown in Fig 2. In this example, each centrality method found a different set of nodes, although all sets were found close to the connection point of the three distinct branches. Interestingly, the seed node is towards the end of a branch and is not at all central. Closeness centrality, as seen in Fig 2B, does the best job at centering the nodes around this connection point, while betweenness and eigenvector select different nodes on branches. Across each graph, we can see nodes highlighted with relevant exceptional attributes. Some attributes, such as the respondent's marital status or primary mode of transportation appeared across multiple types of centrality, while other attributes such as the subject's ethnic background were only found to be exceptional in a single centrality method, as seen in the eigenvector nodes highlighted in Fig 2C.

Complete results for the three largest components within the Chicago referral network are shown in Table 1. Attributes such as 'mstat', which refers to marital status, and 'tmode' which refers to transportation mode, are shared between multiple centrality methods, indicating a higher concentration of individuals with those attributes within the most central nodes of the network. While other attributes, such as 'hisp' and 'raceb', which concern identification as Hispanic or African American, are found to be exceptional within the eigenvector nodes, this observation is not shared between other forms of centrality, indicating this attribute is less likely to be important to the structure of the overall network.

One of the Los Angeles Network components is shown in Fig 3. Here we see a similar structure to Chicago, with several branches extending from a central set of nodes. Unlike Chicago however, we notice significantly more overlap in the sets of nodes produced by each centrality method. This overlap results in a set of attributes identified by each centrality method as

Table 1. Exceptional attributes of Chicago.

Centrality Type	Attribute Code	Attribute Plaintext	Exceptionality
betweenness	mstat-4	Married but currently separated	0.2667
closeness	mstat-5	Currently Divorced	0.3333
	tmode-5	Walks as a primary form of transportation	0.2667
eigenvector	tmode-5	Walks as a primary form of transportation	0.4667
	raceb	Less likely to identify as Black or African American	0.3333
	hisp	More likely to identify as Spanish, Hispanic, or Latino	0.2667

Table 2. Exceptional attributes of Los Angeles.

Centrality Type	Attribute Code	Attribute Plaintext	Exceptionality
betweenness	reside-5	Currently lives in a rented hotel or rooming house	0.3333
	usedi	Has used a drug not otherwise listed on the survey	0.3333
	sexid2-5	Does not have sex only with women	0.3333
	racee	Identifies as a race not otherwise specified by the survey	0.2667
closeness	sexid2-5	Does not have sex only with women	0.4000
	usedi	Has used a drug not otherwise listed on the survey	0.3333
	racee	Identifies as a race not otherwise specified by the survey	0.2667
	reside-5	Currently lives in a rented hotel or rooming house	0.2667
eigenvector	racee	Identifies as a race not otherwise specified by the survey	0.4667
	usedi	Has used a drug not otherwise listed on the survey	0.2667
	reside-5	Currently lives in a rented hotel or rooming house	0.2667

https://doi.org/10.1371/journal.pone.0256601.t002

exceptional, albeit to varying degrees. As shown in Table 2, 'reside-5', indicating living situation, and 'used-i', indicating drug use, have consistent exceptionality scores ranging between 0.267 and 0.333 across each method.

Similar to Chicago, the closeness centrality in Fig 3B provides a comprehensive set of nodes surrounding the center-most node in the component, while in Fig 3A betweenness centrality focuses on the nodes along a path between two branches. The attributes labeled in Fig 3 indicate attributes that are significant to that particular node. These attributes may not necessarily be shared between other central nodes, and therefore may not listed within Table 2. Looking at the prevalence of the attribute 'reside-5', in which the respondent currently rents a room in a hotel or rooming house, one might conclude that central nodes are less likely to have a consistent living situation, a marker for potential homelessness. The attribute 'used-i' indicates that the respondent has used a drug not otherwise listed on the survey. This attribute was also present in each centrality type, indicating higher importance to the overall structure of the network. The attribute 'sexid2-5', which was identified as exceptional by betweenness and closeness centralities indicates that the respondent does not have sexual contact exclusively with women.

Fig 4 shows a single component within the Raleigh-Durham referral network. Like both Chicago and Los Angeles, Fig 4B shows that closeness centrality was effective at identifying the nodes around the center-most point in the component. The nodes identified by betweenness centrality are shown in Fig 4A and eigenvector centrality in Fig 4C.

Key attributes on the component highlighted in Fig 4 include marital status, testing positive for gonorrhea, and having graduated from a 4 year university. When looking at Table 3, we notice a much larger number of attributes marked by high exceptionality. Some attributes, such as 'tmode-5' and 'reside-6' appear across multiple types of centrality, indicating a shared number of central nodes and attributes.

As seen in Table 3, Raleigh-Durham had many more exceptional attributes than other cities. In particular, closeness centrality identified a total of 10 exceptional attributes. The topic of these attributes varies, but several themes seem to appear. Features 'usedc', 'usedh', 'usedb', 'drink1', and 'usedf' relate to substance abuse of heroin, cocaine, sedatives, alcohol, and others. Attributes 'reside-6', 'tmode-5', 'slept-2', and 'slept-3' also appear across multiple centrality types and indicate that the respondent has an inconsistent living situation, ranging from living in a shelter, boarding house, or sleeping in a neighborhood near their home.

Table 3. Exceptional attributes of Raleigh-Durham.

Centrality Type	Attribute Code	Attribute Plaintext	Exceptionality
betweenness	reside-6	Currently lives in a shelter, boarding, or halfway house	0.3333
	slept-2	Slept in a neighborhood, but not a home last week	0.3333
	usedc	Has used heroin and cocaine together	0.2667
closeness	usedc	Has used heroin and cocaine together	0.4000
	usedh	Has used sedatives without a prescription	0.4000
	tmode-5	Walks as a primary form of transportation	0.4000
	usedb	Has used methamphetamines	0.3333
	reside-6	Currently lives in a shelter, boarding, or halfway house	0.3333
	drink1	Has drank alcohol more than average in the past 30 days	0.2667
	slept-3	Has mostly slept in a neighborhood within 20 miles from home	0.2667
	usedf	Has used heroin	0.2667
	slept-2	Slept in a neighborhood, but not a home last week	0.2667
	mstat-5	Currently Divorced	0.2667
eigenvector	tmode-5	Walks as a primary form of transportation	0.4000
	reside-6	Currently lives in a shelter, boarding, or halfway house	0.3333
	rdsq9-1	Has known their recruiter for a very short time (days)	0.3333
	sexid1-9	Does not identify themselves sexually	0.2667
	slept-3	Has mostly slept in a neighborhood within 20 miles from home	0.2667
	slept-2	Slept in a neighborhood, but not a home last week	0.2667

Discussion

In most cases, even seemingly minor attributes identified by exceptional centrality values can be linked back to issues associated either with living with HIV or with the transmission of HIV. Several of the most important examples are discussed below. These examples show that our centrality selection method is able to confirm existing research in a number of cases, as well as to identify new attributes for intervention.

In both Chicago and Raleigh-Durham, martial separation and divorce were found at high rates among nodes with high betweenness and closeness centrality. The relationship between divorce and HIV has been studied in some contexts, such as [47], in which it is is speculated that divorce can be used as a protective mechanism against a spouse who engages in risky sexual behavior, and also that divorce rates should increase in response to risky behaviors in populations with increasing rates of HIV. Both of these situations are true of the Chicago population. Other studies have found high rates of divorce among the HIV population generally [48, 49] and HIV-discordant couples specifically [50]. This is an area where interventions by social workers providing help for couples living with HIV would provide great benefit.

Eigenvector centrality identified several members of a Hispanic community in Chicago based on the their connections to each other. Hispanic communities have been extensively studied in relation to HIV. In particular, it has been speculated that strong social networks have been underused as a resource for spreading prevention information [51], for increasing individual participation in testing [52], and for decreasing risky sexual behaviors [53]. A classic RDS study of Latino men by Ramierez [54] demonstrates that self-protection by Latino men is increased with participation in community involvement activities.

An individual high centrality node in Raleigh reported testing positive for gonorrhea. New HIV diagnoses have been associated with gonorrhea diagnoses [55], particularly among MSM

[56]. On the other hand, PrEP has resulted in an outbreak of STDs independent of HIV, including chlamydia [57] and syphillis [58] in addition to gonorrhea.

In Raleigh, the attribute of 'Age of first sex lower than average' was detected by eigenvector centrality. In particular, if early sex was the result of sexual abuse, associations have been made with HIV. Child sexual abuse has been shown to lead to risky sexual behavior in both heterosexual men [59] and MSM [60], and is associated with a higher probabilty of being HIV-positive [61].

In Chicago and Raleigh-Durham, participants who walk as a primary form of transportation show exceptional centrality scores. There is research showing that transportation vulnerability is a major issue for HIV-susceptible populations. In particular, transportation vulnerable HIV-positive individuals do not have easy access to care services [62] and are less likely to maintain viral suppression [63]. It has been suggested that providing transportation assistance to susceptible populations decreases the risk of contracting HIV [64]. Our research lends support to this idea, and suggests that it should be pursued further.

Across all three cities, we find that two major themes appear. Firstly, central nodes appear to have higher rates of inconsistent living situations. Attributes such as 'reside-5' which describes respondents living in rented hotels or rooming homes, and 'slept-3' which describes respondents sleeping in a neighborhood within 20 miles of home are both indicators of higher risks of homelessness in these populations. Studies have found that HIV-positive homeless people had more sex partners, were more likely to exchange sex for money or drugs, and had more unprotected sex than those with homes [65]. Unstable housing has been associated with an increased risk of acquiring HIV for youth [66], and in [67] it is stated that the HIV-positive homeless are 3.84 times more likely to have incomplete viral suppression than the housed. In a San Francisco study, homelessness at diagnosis of HIV was associated with a higher probability of death [68]. The strength of the centrality results for this attribute, combined with the seriousness of the related research, suggest that any interventions by government or charities to reduce homelessness will have an impact on the spread of HIV.

Secondly, nodes identified by the centrality algorithms were more prone to substance abuse. Attributes such as 'usedi' which indicated the respondent used a drug not otherwise listed, and 'usedc' which indicated the user had used heroin and cocaine together are examples of the several attributes identified by the centrality algorithms that describe the respondent's substance abuse. Drug use has been suggested as a factor in HIV transmission, although there is debate as to whether sexual or parenteral risk is greater for PWIDs [69]. Other evidence suggests drug use as a causal risk factor for HIV [70]. Due to the connection between drug use and HIV transmission, drug use treatment has been proposed as an HIV prevention strategy [71]. This research supports that proposition.

Conclusion

This paper discussed the use of centrality based network algorithms to identify biomarkers for the spread of HIV in high risk communities. We utilized the SATHCAP dataset to create a network of referrals across three US cities based upon sexual activity and drug use—behaviors known to transmit HIV. To determine the central, important nodes in these networks, we calculated three forms of network centrality: betweenness, closeness, and eigenvector. By identifying the key attributes of the discovered central nodes, we provide a clear aim for targeted intervention campaigns to both prevent HIV infection and improve the lives of those infected.

The methodology was shown to be successful both in finding established biomarkers, and in identifying new areas for further research. Across the central nodes of the SATHCAP referral network, we found minor themes such as divorce and diagnosis of other STDs. Major

themes were also found, such as uncertainty in living situation and substance use. Attributes with high occurrence and relatively small amounts of related research, such as transportation vulnerability, were seen as potential areas for further investigation.

There are many RDS surveys devoted to obtaining biological and behavioral information about HIV populations [72]. Conditions in the world and related behaviors change quickly [73, 74]. This methodology for creating networks from RDS survey data is not exclusive to the SATHCAP dataset. By creating networks where respondents are represented as nodes and recruitments as edges, this methodology can be duplicated on similar datasets. Nodes identified as central to the structure of the RDS network are more likely to have a pronounced impact on the spread of HIV, and make good targets for focused intervention campaigns. Our research found several relevant attributes, and analyses of additional surveys are likely find more.

Author Contributions

Conceptualization: Jacob Grubb, John Matta.

Data curation: Jacob Grubb, Derek Lopez, Bhuvaneshwar Mohan.

Formal analysis: Jacob Grubb.

Methodology: Jacob Grubb.

Project administration: Jacob Grubb.

Software: Jacob Grubb.

Supervision: John Matta.

Visualization: Jacob Grubb.

Writing – original draft: Jacob Grubb, John Matta.

Writing – review & editing: Jacob Grubb, John Matta.

References

1. Eames KT. Modelling disease spread through random and regular contacts in clustered populations. Theoretical population biology. 2008; 73(1):104–111. https://doi.org/10.1016/j.tpb.2007.09.007 PMID: 18006032

2. Lloyd AL, May RM. How viruses spread among computers and people. Science. 2001; 292 (5520):1316–1317. https://doi.org/10.1126/science.1061076 PMID: 11360990

3. Linley L, Johnson AS, Song R, Wu B, Hu S, Singh S, et al. Estimated HIV incidence and prevalence in the United States 2010–2015. 2018.

4. Hess KL, Johnson SD, Hu X, Li J, Wu B, Yu C, et al. Diagnoses of HIV infection in the United States and dependent areas, 2017. 2018.

5. Pellowski JA, Kalichman SC, Matthews KA, Adler N. A pandemic of the poor: social disadvantage and the US HIV epidemic. American Psychologist. 2013; 68(4):197.

6. Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. Social problems. 1997; 44(2):174–199. https://doi.org/10.1525/sp.1997.44.2.03x0221m

7. Compton W, Normand J, Lambert E. Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP), July 2009. Journal of Urban Health. 2009; 86(1):1–4. https://doi.org/10.1007/s11524-009-9373-4 PMID: 19543977

8. Lee S, Suzer-Gurtekin T, Wagner J, Valliant R. Total survey error and respondent driven sampling: focus on nonresponse and measurement errors in the recruitment process and the network size reports and implications for inferences. Journal of Official Statistics. 2017; 33(2):335–366. https://doi.org/10.1515/jos-2017-0017

9. Bearman PS, Moody J, Stovel K. Chains of affection: The structure of adolescent romantic and sexual networks. American journal of sociology. 2004; 110(1):44–91. https://doi.org/10.1086/386272

10. Potterat JJ, Phillips-Plummer L, Muth SQ, Rothenberg R, Woodhouse D, Maldonado-Long T, et al. Risk network structure in the early epidemic phase of HIV transmission in Colorado Springs. Sexually transmitted infections. 2002; 78(suppl 1):i159–i163. https://doi.org/10.1136/sti.78.suppl_1.i159 PMID: 12083437

11. Grubb J, Lopez D, Mohan B, Matta J. Identifying Biomarkers for Important Nodes in Networks of Sexual and Drug Activity. In: International Conference on Complex Networks and Their Applications. Springer; 2020. p. 357–369.

12. Liljeros F, Edling CR, Amaral LAN, Stanley HE, Åberg Y. The web of human sexual contacts. Nature. 2001; 411(6840):907–908. https://doi.org/10.1038/35082140 PMID: 11418846

13. Crepaz N, Hess KL, Purcell DW, Hall HI. Estimating national rates of HIV infection among MSM, persons who inject drugs, and heterosexuals in the United States. Aids. 2019; 33(4):701–708. https://doi.org/10.1097/QAD.0000000000002111 PMID: 30585840

14. Rodger AJ, Cambiano V, Bruun T, Vernazza P, Collins S, Van Lunzen J, et al. Sexual activity without condoms and risk of HIV transmission in serodifferent couples when the HIV-positive partner is using suppressive antiretroviral therapy. Jama. 2016; 316(2):171–181. https://doi.org/10.1001/jama.2016.5148 PMID: 27404185

15. Gantenberg JR, King M, Montgomery MC, Galárraga O, Prosperi M, Chan PA, et al. Improving the impact of HIV pre-exposure prophylaxis implementation in small urban centers among men who have sex with men: an agent-based modelling study. PloS one. 2018; 13(7):e0199915. https://doi.org/10.1371/journal.pone.0199915 PMID: 29985949

16. Youm Y, Mackesy-Amiti ME, Williams CT, Ouellet LJ. Identifying hidden sexual bridging communities in Chicago. Journal of Urban Health. 2009; 86(1):107–120. https://doi.org/10.1007/s11524-009-9371-6 PMID: 19543836

17. Ober A, Shoptaw S, Wang PC, Gorbach P, Weiss RE. Factors associated with event-level stimulant use during sex in a sample of older, low-income men who have sex with men in Los Angeles. Drug and Alcohol Dependence. 2009; 102(1-3):123–129. https://doi.org/10.1016/j.drugalcdep.2009.02.002 PMID: 19327917

18. Firestone SM, Ward MP, Christley RM, Dhand NK. The importance of location in contact networks: Describing early epidemic spread using spatial social network analysis. Preventive veterinary medicine. 2011; 102(3):185–195. https://doi.org/10.1016/j.prevetmed.2011.07.006 PMID: 21852007

19. Brooks D, Matta J. Spreading of rodent infestations through a city. Applied Network Science. 2019; 4 (1):89. https://doi.org/10.1007/s41109-019-0212-6

20. De P, Singh AE, Wong T, Yacoub W, Jolly A. Sexual network analysis of a gonorrhoea outbreak. Sexually transmitted infections. 2004; 80(4):280–285. https://doi.org/10.1136/sti.2003.007187 PMID: 15295126

21. Wylie JL, Jolly A. Patterns of chlamydia and gonorrhea infection in sexual networks in Manitoba, Canada. Sexually transmitted diseases. 2001; 28(1):14–24. https://doi.org/10.1097/00007435-200101000-00005 PMID: 11196040

22. Oster AM, Wejnert C, Mena LA, Elmore K, Fisher H, Heffelfinger JD. Network analysis among HIV-infected young black men who have sex with men demonstrates high connectedness around few venues. Sexually transmitted diseases. 2013; 40(3):206. https://doi.org/10.1097/OLQ.0b013e3182840373 PMID: 23403601

23. Liljeros F, Edling CR, Amaral LAN. Sexual networks: implications for the transmission of sexually transmitted infections. Microbes and infection. 2003; 5(2):189–196. https://doi.org/10.1016/S1286-4579(02)00058-8 PMID: 12650777

24. Anderson R, Gupta S, Ng W, et al. The significance of sexual partner contact networks for the transmission dynamics of HIV. J Acquir Immune Defic Syndr. 1990; 3(4):417–429. PMID: 2179528

25. Crawford FW. The graphical structure of respondent-driven sampling. Sociological methodology. 2016; 46(1):187–211. https://doi.org/10.1177/0081175016641713 PMID: 31607761

26. Holme P, Kim BJ, Yoon CN, Han SK. Attack vulnerability of complex networks. Phys Review. 2002; 65 (5):056109. PMID: 12059649

27. Bell DC, Atkinson JS, Carlson JW. Centrality measures for disease transmission networks. Social networks. 1999; 21(1):1–21. https://doi.org/10.1016/S0378-8733(98)00010-0

28. Büttner K, Krieter J, Traulsen A, Traulsen I. Static network analysis of a pork supply chain in Northern Germany—Characterisation of the potential spread of infectious diseases via animal movements. Preventive veterinary medicine. 2013; 110(3-4):418–428. https://doi.org/10.1016/j.prevetmed.2013.01.008 PMID: 23462679

29. Ribeiro SP, e Silva AC, Dáttilo W, Reis AB, Góes-Neto A, Alcantara LCJ, et al. Severe airport sanitarian control could slow down the spreading of COVID-19 pandemics in Brazil. PeerJ. 2020; 8:e9446. https://doi.org/10.7717/peerj.9446 PMID: 32617196

30. Carreras I, Miorandi D, Canright GS, Engo-Monsen K. Understanding the spread of epidemics in highly partitioned mobile networks. In: Proceedings of the 1st international conference on Bio inspired models of network, information and computing systems; 2006. p. 2–es.

31. Ping L, Jie Z, Xiao-Ke X, Michael S. Dynamical influence of nodes revisited: A markov chain analysis of epidemic process on networks. Chinese Physics Letters. 2012; 29(4):048903. https://doi.org/10.1088/0256-307X/29/4/048903

32. Basnarkov L. SEAIR Epidemic spreading model of COVID-19. Chaos, Solitons & Fractals. 2021; 142:110394. https://doi.org/10.1016/j.chaos.2020.110394 PMID: 33162690

33. Natale F, Giovannini A, Savini L, Palma D, Possenti L, Fiore G, et al. Network analysis of Italian cattle trade patterns and evaluation of risks for potential disease spread. Preventive veterinary medicine. 2009; 92(4):341–350. https://doi.org/10.1016/j.prevetmed.2009.08.026 PMID: 19775765

34. Binnewijzend MA, Adriaanse SM, Van der Flier WM, Teunissen CE, de Munck JC, Stam CJ, et al. Brain network alterations in Alzheimer's disease measured by eigenvector centrality in fMRI are related to cognition and CSF biomarkers. Human brain mapping. 2014; 35(5):2383–2393. https://doi.org/10.1002/hbm.22335 PMID: 24039033

35. Heckathorn DD. Comment: Snowball versus respondent-driven sampling. Sociological methodology. 2011; 41(1):355–366. https://doi.org/10.1111/j.1467-9531.2011.01244.x PMID: 22228916

36. Handcock MS, Gile KJ, Mar CM. Estimating hidden population size using respondent-driven sampling data. Electronic journal of statistics. 2014; 8(1):1491. https://doi.org/10.1214/14-EJS923 PMID: 26180577

37. Barabási AL, Bonabeau E. Scale-free networks. Scientific american. 2003; 288(5):60–69. https://doi.org/10.1038/scientificamerican0503-60 PMID: 12701331

38. Barabási AL. Network science. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2013; 371(1987):20120375. https://doi.org/10.1098/rsta.2012.0375 PMID: 23419844

39. Lee SH, Kim PJ, Jeong H. Statistical properties of sampled networks. Physical review E. 2006; 73(1):016102. https://doi.org/10.1103/PhysRevE.73.016102 PMID: 16486211

40. Wejnert C. 3. An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and Out-of-Equilibrium Data. Sociological methodology. 2009; 39(1):73–116. PMID: 20161130

41. Chen S, Lu X. An immunization strategy for hidden populations. Scientific reports. 2017; 7(1):1–10. https://doi.org/10.1038/s41598-017-03379-4 PMID: 28607411

42. Maharani W, Gozali AA, et al. Degree centrality and eigenvector centrality in twitter. In: 2014 8th international conference on telecommunication systems services and applications (TSSA). IEEE; 2014. p. 1–5.

43. Umadevi V. Case study–centrality measure analysis on co-authorship network. Journal of Global Research in Computer Science. 2013; 4(1):67–70.

44. Freeman LC. The gatekeeper, pair-dependency and structural centrality. Quality and Quantity. 1980; 14(4):585–592. https://doi.org/10.1007/BF00184720

45. Matta J, Ercal G, Sinha K. Comparing the speed and accuracy of approaches to betweenness centrality approximation. Computational Social Networks. 2019; 6(1):2. https://doi.org/10.1186/s40649-019-0062-5

46. Zhang J, Luo Y. Degree centrality, betweenness centrality, and closeness centrality in social network. In: Proceedings of the 2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017). vol. 132; 2017. p. 300–303.

47. Reniers G. Marital strategies for regulating exposure to HIV. Demography. 2008; 45(2):417–438. https://doi.org/10.1353/dem.0.0002 PMID: 18613488

48. Barkan SE, Melnick SL, Preston-Martin S, Weber K, Kalish LA, Miotti P, et al. The women's interagency HIV study. Epidemiology. 1998; p. 117–125. https://doi.org/10.1097/00001648-199803000-00004 PMID: 9504278

49. Brown MJ, Cohen SA, DeShazo JP. Psychopathology and HIV diagnosis among older adults in the United States: disparities by age, sex, and race/ethnicity. Aging & mental health. 2020; 24(10):1746–1753. https://doi.org/10.1080/13607863.2019.1636201

50. Mackelprang RD, Bosire R, Guthrie BL, Choi RY, Liu A, Gatuguta A, et al. High rates of relationship dissolution among heterosexual HIV-serodiscordant couples in Kenya. AIDS and Behavior. 2014; 18(1):189–193. https://doi.org/10.1007/s10461-013-0529-6 PMID: 23728523

51. Painter TM. Social support networks: An underutilized resource for the prevention of HIV and other sexually transmitted diseases among Hispanic/Latino migrants and immigrants. Journal of health care for the poor and underserved. 2018; 29(1):44–57. https://doi.org/10.1353/hpu.2018.0005 PMID: 29503286

52. Painter TM, Song EY, Mullins MM, Mann-Jackson L, Alonzo J, Reboussin BA, et al. Social support and other factors associated with HIV testing by Hispanic/Latino gay, bisexual, and other men who have sex with men in the US South. AIDS and Behavior. 2019; 23(3):251–265. https://doi.org/10.1007/s10461-019-02540-6 PMID: 31102108

53. Althoff MD, Theall K, Schmidt N, Hembling J, Gebrekristos HT, Thompson MM, et al. Social support networks and HIV/STI risk behaviors among Latino immigrants in a new receiving environment. AIDS and Behavior. 2017; 21(12):3607–3617. https://doi.org/10.1007/s10461-017-1849-8 PMID: 28733921

54. Ramirez-Valles J. The protective effects of community involvement for HIV risk behavior: a conceptual framework. Health Education Research. 2002; 17(4):389–403. https://doi.org/10.1093/her/17.4.389 PMID: 12197585

55. Barbee LA, Khosropour CM, Dombrowksi JC, Golden MR. New HIV diagnosis independently associated with rectal gonorrhea and chlamydia in men who have sex with men. Sexually transmitted diseases. 2017; 44(7):385. https://doi.org/10.1097/OLQ.0000000000000614 PMID: 28608786

56. Jones J, Weiss K, Mermin J, Dietz P, Rosenberg ES, Gift TL, et al. Proportion of incident HIV cases among men who have sex with men attributable to gonorrhea and chlamydia: a modeling analysis. Sexually transmitted diseases. 2019; 46(6):357. https://doi.org/10.1097/OLQ.0000000000000980 PMID: 31095100

57. Biello KB, Edeza A, Montgomery MC, Almonte A, Chan PA. Risk perception and interest in HIV pre-exposure prophylaxis among men who have sex with men with rectal gonorrhea and chlamydia infection. Archives of sexual behavior. 2019; 48(4):1185–1190. https://doi.org/10.1007/s10508-018-1260-1 PMID: 30105617

58. Fujimoto K, Flash CA, Kuhns LM, Kim JY, Schneider JA. Social networks as drivers of syphilis and HIV infection among young men who have sex with men. Sexually transmitted infections. 2018; 94(5):365–371. https://doi.org/10.1136/sextrans-2017-053288 PMID: 29440465

59. Peterson ZD, Janssen E, Goodrich D, Fortenberry JD, Hensel DJ, Heiman JR. Child sexual abuse and negative affect as shared risk factors for sexual aggression and sexual HIV risk behavior in heterosexual men. Archives of sexual behavior. 2018; 47(2):465–480. https://doi.org/10.1007/s10508-017-1079-1 PMID: 29090393

60. Batchelder AW, Ehlinger PP, Boroughs MS, Shipherd JC, Safren SA, Ironson GH, et al. Psychological and behavioral moderators of the relationship between trauma severity and HIV transmission risk behavior among MSM with a history of childhood sexual abuse. Journal of behavioral medicine. 2017; 40(5):794–802. https://doi.org/10.1007/s10865-017-9848-9 PMID: 28396969

61. Wu E. Childhood sexual abuse among Black men who have sex with men: A cornerstone of a syndemic? PLoS one. 2018; 13(11):e0206746. https://doi.org/10.1371/journal.pone.0206746 PMID: 30383859

62. Sagrestano LM, Clay J, Finerman R, Gooch J, Rapino M. Transportation vulnerability as a barrier to service utilization for HIV-positive individuals. Aids Care. 2014; 26(3):314–319. https://doi.org/10.1080/09540121.2013.819403 PMID: 23876086

63. Goswami ND, Schmitz MM, Sanchez T, Dasgupta S, Sullivan P, Cooper H, et al. Understanding local spatial variation along the care continuum: the potential impact of transportation vulnerability on HIV linkage to care and viral suppression in high-poverty areas, Atlanta, Georgia. Journal of acquired immune deficiency syndromes (1999). 2016; 72(1):65. https://doi.org/10.1097/QAI.0000000000000914 PMID: 26630673

64. Swann M. Economic strengthening for HIV prevention and risk reduction: a review of the evidence. AIDS Care. 2018; 30(sup3):37–84. https://doi.org/10.1080/09540121.2018.1479029

65. Kidder DP, Wolitski RJ, Pals SL, Campsmith ML. Housing status and HIV risk behaviors among homeless and housed persons with HIV. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2008; 49(4):451–455. https://doi.org/10.1097/QAI.0b013e31818a652c PMID: 19186357

66. Marshall BD, Kerr T, Shoveller JA, Patterson TL, Buxton JA, Wood E. Homelessness and unstable housing associated with an increased risk of HIV and STI transmission among street-involved youth. Health & place. 2009; 15(3):783–790. https://doi.org/10.1016/j.healthplace.2008.12.005 PMID: 19201642

67. Thakarar K, Morgan JR, Gaeta JM, Hohl C, Drainoni ML. Homelessness, HIV, and incomplete viral suppression. Journal of health care for the poor and underserved. 2016; 27(1):145. https://doi.org/10.1353/hpu.2016.0020 PMID: 27528794

68. Spinelli MA, Hessol NA, Schwarcz S, Hsu L, Parisi MK, Pipkin S, et al. Homelessness at diagnosis is associated with death among people with HIV in a population-based study of a US city. AIDS (London, England). 2019; 33(11):1789. https://doi.org/10.1097/QAD.0000000000002287

69. Strathdee SA, Sherman SG. The role of sexual transmission of HIV infection among injection and non-injection drug users. Journal of Urban Health. 2003; 80(3):iii7–iii14. https://doi.org/10.1093/jurban/jtg078 PMID: 14713667

70. Drumright LN, Patterson TL, Strathdee SA. Club drugs as causal risk factors for HIV acquisition among men who have sex with men: a review. Substance use & misuse. 2006; 41(10-12):1551–1601. https://doi.org/10.1080/10826080600847894 PMID: 17002993

71. Metzger DS, Navaline H, Woody GE. Drug abuse treatment as AIDS prevention. Public Health Reports. 1998; 113(Suppl 1):97. PMID: 9722815

72. Malekinejad M, Johnston LG, Kendall C, Kerr LRFS, Rifkin MR, Rutherford GW. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. AIDS and Behavior. 2008; 12(1):105–130. https://doi.org/10.1007/s10461-008-9421-1 PMID: 18561018

73. Guest G, Shattuck D, Johnson L, Akumatey B, Clarke EEK, Chen PL, et al. Changes in sexual risk behavior among participants in a PrEP HIV prevention trial. Sexually transmitted diseases. 2008; 35 (12):1002–1008. https://doi.org/10.1097/OLQ.0b013e3181812939 PMID: 19051397

74. Buffie WC. Public health implications of same-sex marriage. American Journal of Public Health. 2011; 101(6):986–990. https://doi.org/10.2105/AJPH.2010.300112 PMID: 21493934