# Inferring the organizing principles of young men who have sex with men from combining many networks

Balint Neray, PhD[1]; Johan Koskinen, PhD[2]; Patrick Janulis, PhD[1]; Gregory Phillips II, PhD[1]; Michelle Birkett, PhD[1]

[1] Department of Medical Social Sciences, Feinberg School of Medicine; Institute for Sexual and Gender Minority Health and Wellbeing; Northwestern University
[2] The Mitchell Centre for SNA; University of Manchester

**Northwestern**
Institute for Sexual and Gender Minority Health and Wellbeing

**Northwestern Medicine®**
Feinberg School of Medicine

## INTRODUCTION

The sexual network topology of individuals and their communities may affect the spread of infectious diseases, both within and across specific populations. To better understand this process, we collected ego-centric data on the personal sexual networks of 175 young men who have sex with men (YMSM). Ego-centric data, however, are incomplete, which makes studying disease spread difficult (Krivitsky and Morris, 2017; Rolls et al., 2013; Handcock and Gile, 2010).

For this reason, a semi-supervised entity-resolution scheme was designed in order to identify unique individuals and match their attribute data. To account for lack of complete information on alter-alter ties, we fit an exponential random graph model (ERGM; Lusher et al., 2013) using a Bayesian data augmentation algorithm (Koskinen et al., 2013). The algorithm also provides us with a distribution of networks with imputed ties, allowing for further investigation of structures, such as clustering (triad closure), that are crucial for studying disease spread.

## DESIGN/METHODOLOGY

Enrollment criteria for the parent study included English-speaking, assigned a male sex at birth, had a prior sexual encounter with a male or identified as gay or bisexual, resided in the metropolitan area of Chicago, and were available for multiple follow-up visits in the subsequent 24 months. Respondents were asked to nominate sexual contacts and the sexual contacts of their nominees (n=602). Network member nomination was capped at 40, although a minority of participants reached the cap. Among the 741 male network members reported, 360 were Black, 164 Latino, 156 White, and 61 other race/ethnicity. Most (82.9%) identified their sexual orientation identity as completely or mostly gay.

For the ERGM, we used a standard specification for a Markov and Social Circuit dependence model (Snijders et al., 2006) with structural effects (density, degree, and triad closure), main covariate effects (age and race/ethnicity), homophily terms (on race/ethnicity and age), and one dyadic covariate (geographical distance). It is assumed that the likelihood to observe a tie can be calculated by a collection of graph statistics derived from dependence assumptions for the ties (Frank and Strauss, 1986; Snijders et al., 2006 ), and a vector of other statistical parameters representing individual and dyadic attributes.

Figure 1. Observed network (on the left), and an example of an imputed network (on the right)
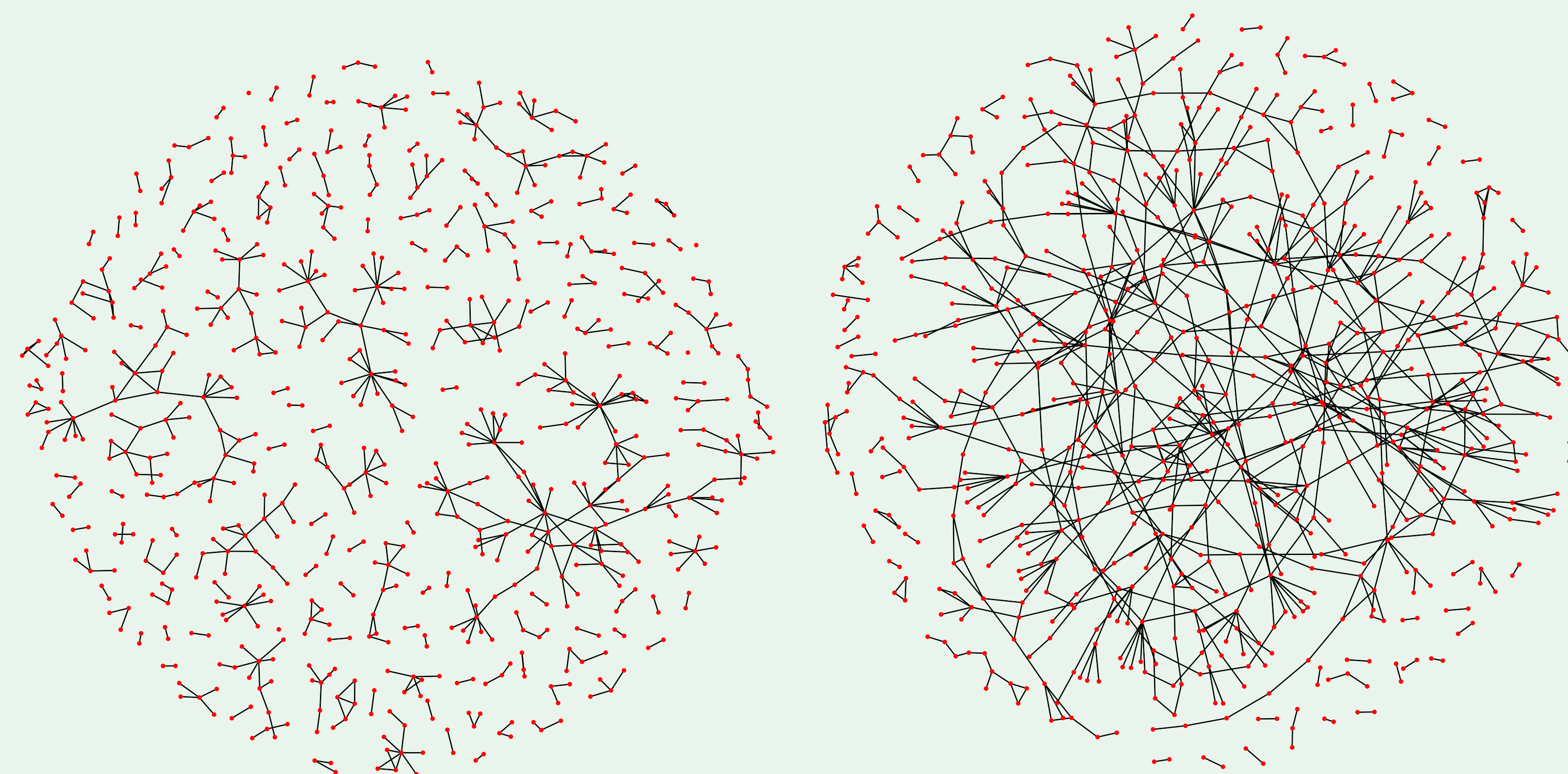


## KEY FINDINGS

### Network reconstruction

Entity resolution yielded a reconstructed network with 628 observed sexual ties among 139 egos and 602 alters, missing ties were imputed using the Bayesian data augmentation algorithm. The degree distribution of the empirical network is almost identical to the degree distribution of the simulated networks, which indicates that the augmented data provide a good representation of the empirically observed data.

### Network analysis

A number of hierarchically nested models were run and Figure 2 provides preliminary results from the final model. This YMSM sexual network is strongly shaped by racial and ethnic homophily, as the likelihood of observing a Black-Black or a Latino-Latino tie is even higher than observing a White-White tie. Furthermore, sex ties are more common between individuals who live closer to each other and are more similar in age. The degree variable takes into account that certain individuals have a preference for having a larger number of sex partners. In addition, the triadic closure variable accounts for the tendency of groups of individuals to cluster, that is to have sex with each other. Most importantly, we found lower tendency for clustering among Black individuals, which is indirect evidence for longer reach in their network. This result holds crucial implications for understanding racial disparities in HIV prevalence.

Figure 2. Posterior distributions for ERGM parameters
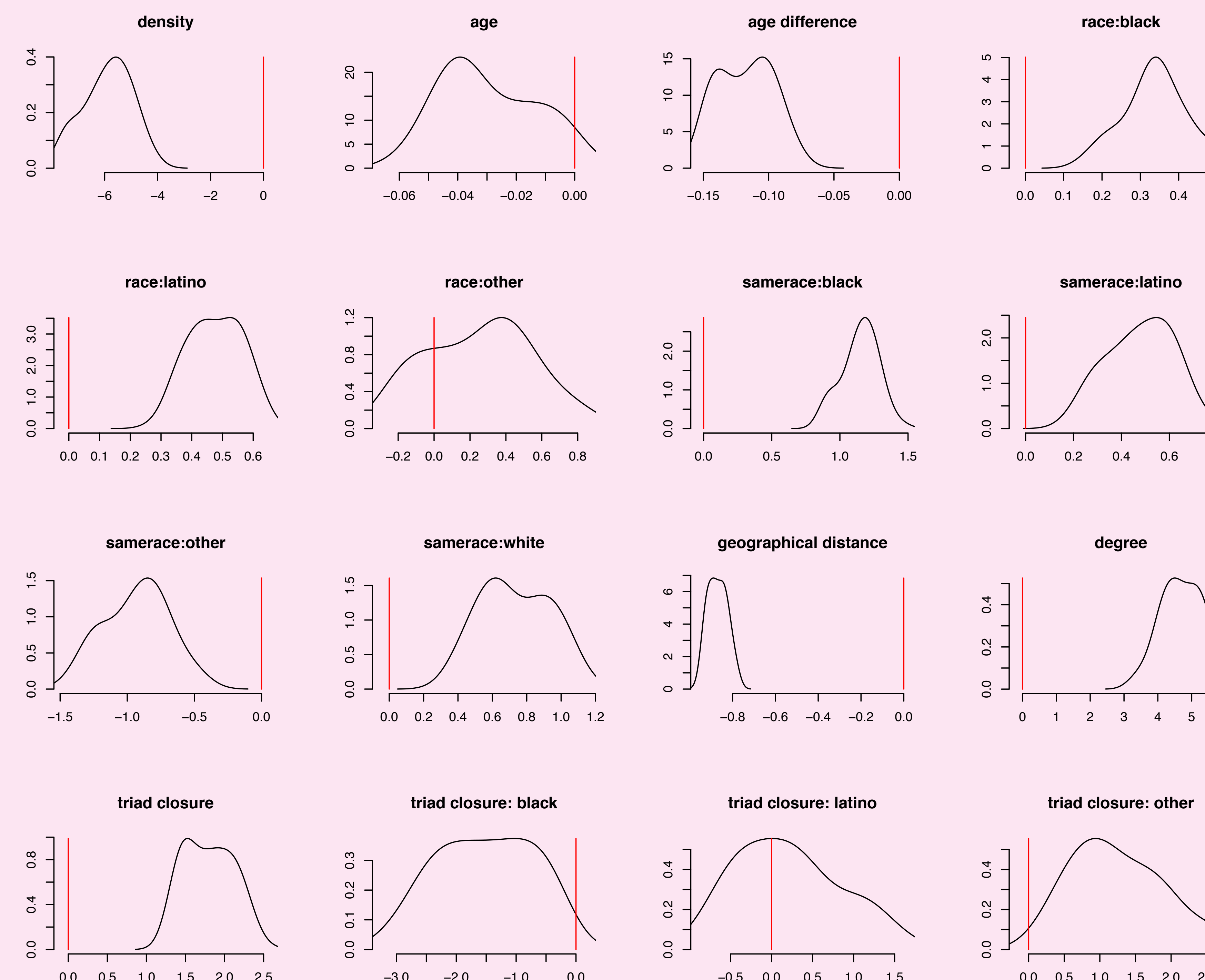


## SUMMARY AND IMPLICATIONS

- YMSM are considered to be a "hard to reach" population; as such data collection is difficult and often incomplete.

- We are in the process of developing solutions to work with such data by utilizing an entity-resolution scheme and a Bayesian augmentation algorithm for network data.

- From a substantive point of view, our work contributes to the research on racial disparities in HIV prevalence.

- Greater HIV prevalence among Black individuals, particularly MSM, is a well documented phenomenon. However, research shows they engage in fewer risk behaviors than other MSM groups.

- Earlier research shows that the strongly racially segregated structure of YMSM sexual networks helps maintain unequal HIV prevalence across groups.

- Our work adds to this argument by providing preliminary evidence for lower clustering and consequently longer reach within the Black network.

- Since infectious diseases spread more efficiently in such networks, these results suggest that the difference in the network structure of Black YMSM may be responsible for higher HIV incidence and prevalence within this group.

## NEXT STEPS / IMPLICATIONS

- While developing better data collection procedures and tools for hard to reach populations is crucial, learning how to deal with messy and incomplete data is just as important.

- Our work is limited by the semi-supervised nature of the entity-resolution scheme, but our team is working on automated solutions.

- The Bayesian augmentation algorithm can get computationally very intensive; hence, further work is needed to make this process more efficient.

- The results that include heterogeneous triangles are preliminary as they based on very few (400) iterations. Further samples are pending.

- If these results hold, we have made a significant step towards better understanding racial disparities in HIV prevalence.