-
- The Challenges of Surveying Heavy-tail Distributions for Use in
- Infectious Disease Dynamics
-
- Michael E. DeWitt^{1,2} Nicholas Kortessis² John W. Sanders¹ Candice J. McNeil¹
- 1 Section on Infectious Diseases, Wake Forest University School of Medicine,
- Winston-Salem, NC, USA
- 2 Department of Biology, Wake Forest University, Winston-Salem, NC, USA
- Corresponding author: medewitt@wakehealth.edu

Abstract

 Sexual networks often have heavy-tails, where a small number of exceptional individuals in a population have many more sexual partners than the average (e.g., more than five standard deviations). Heavy-tails pose challenges when surveying this group, as these exceptional in- dividuals are uncommon in the population (and so hard to detect), but have disproportionate impact on epidemiological questions, such as those related to the spread of sexually transmit- ted diseases. In essence, omitting these individuals is a severe error. In this modeling study, we use prior estimates of the distribution of sexual partners amongst men who have sex with men to explore the implication of different sample sizes on survey estimates. We find that even large surveys consistently fail to capture the variance of the sexual network. Surveys of heavy-tailed sexual networks should be designed with this high variance in mind so as not to underestimate the disease dynamics. The failure to adequately capture the variance within a heavy-tailed network has strong implications for infectious disease dynamics and modeling as disease dynamics are often driven by the heavy-tail.

Introduction

 In early 2022, mpox (formerly known as monkeypox) emerged out of traditionally endemic areas of eastern and central Africa and was identified across the world [\[1\]](#page-15-0). Monkeypox virus, the virus which causes mpox, was traditionally thought to have lower epidemic potential due to low rates of human-to-human transmission [\[2\]](#page-15-1). Human-to-human transmission typically occurs by close and prolonged skin-to-skin contact, in addition to fomites and respiratory droplets [\[3\]](#page-15-2). Prior epidemiological analysis and modeling studies had estimated basic repro- duction numbers less than 1, indicating low potential for epidemic growth [\[2,](#page-15-1) [4\]](#page-15-3). However, many gay, bisexual, and other men who have sex with men were being diagnosed after re- ported sexual contact [\[5\]](#page-15-4), suggesting sustained mpox spread. Mathematical modeling studies were conducted to explain mpox spread despite the thought that the disease should stutter to a stop. These studies indicated that the highly diffuse sexual networks among men who have sex with men may allow for the virus to achieve a basic reproduction number greater than one [\[6\]](#page-15-5). Furthermore, Endo and colleagues have referred to their findings as evidence of "heavy-tailed sexual contact networks", a concept related to probability distributions in ³⁹ which probability density for larger values is higher than that of the exponential distribution. This phenomena relates to the superspreader phenomena which has been observed in other infectious disease processes where a few infections generate a large number of infections due $\frac{42}{1}$ to large variations in individual infectiousness [\[7,](#page-16-0) [8\]](#page-16-1). In the context of sexual networks, there is extremely high variability between reported sexual partners with a very small proportion of individuals having accounting for most sexual partners (Fig. [1\)](#page-6-0). These few, exceptional

 individuals, greatly increase the population-level basic reproductive number. The outbreak of mpox has shown that understanding the role of these sexual networks is vital in modeling the potential for infectious disease transmission.

⁴⁸ The role of sexual networks on the transmission dynamics of infections and sexually transmitted transmitted infections more specifically has been well understood $[9-12]$ $[9-12]$. As more contacts are made, there are more opportunities for infected hosts to transmit the pathogen to their contacts. These insights have been integrated into many mathematical models estimating how infections move within sexual networks [\[13,](#page-17-1) [14\]](#page-17-2). However, many of the data used to generate the network models are limited in the number of participants due to the private nature of the data and the vast heterogeneity in sexual practices. As shown by Endo et al., the heavy-tail of the number of partners complicates understanding these unseen networks [\[6\]](#page-15-5). Furthermore, the role that sample size plays in appropriately capturing the heterogeneity of these heavy-tailed networks remains largely unexplored. Achieving proper estimates for modeling sexual networks is vital in developing strategies for disease control ₅₉ and outbreak prevention. In this paper we explore the implications of different sample sizes given existing literature on sexual networks.

Material and Methods

Data sources

 Following the work of Endo et al. we examine the truncated discrete Weibull distributions which were fitted to empirical data from sexual partnerships from the National Surveys of

 Sexual Attitudes and Lifestyles (NATSAL) from three survey waves: Natsal-2 (1999–2000; 12,110 participants) (26), Natsal-3 (2010–2012; 15,162 participants) [\[6,](#page-15-5) [15\]](#page-17-3). The NATSAL data includes information on the reported number of sexual partners over a given year for those aged 18-44. In this analysis we considered only the distribution of partners amongst men who have sex with other men (MSM) which were estimated from 409 participants in the combined data sets. The Weibull distribution, a continuous probability distribution π with support for positive real numbers, has two parameters: a scale parameter α and shape τ_1 parameter β . The shape of the distribution is primarily governed by the scale parameter with values less than one resulting in more mass towards zero while scale values greater than one have more mass in the tails. However, the mass in the tails is greater than that of the exponential distribution. When the scale parameter is one, the Weibull distribution reduces to the exponential distribution, a distribution which is commonly used to model π the number of contacts in social networks. In order to account for the discrete nature of sexual partners, we passed the continuous estimates generated from the Weibull distribution ₇₉ through a Poisson distribution in order to account for the discrete nature of sexual contacts. We utilized the parameter estimates for MSM sexual networks from Endo et al. in (Table [1\)](#page-5-0). These parameters result in a mean number of annual sexual partners of 9.24 amongst 82 MSM, however, nearly 70% of the population has fewer than 5 partners, but nearly 5% have more than 30 (Fig. [1\)](#page-6-0). The estimates for this distribution are broadly consistent with the ⁸⁴ findings from the 4,904 participant study by Weiss and colleagues who found a mean number of annual active partners of 8.5 [\[13\]](#page-17-1).

Table 1: Weibull distribution parameters estimates from Endo et al.

Estimation of survey representativeness

⁸⁷ To understand the role of sample size on estimates of the number of sexual partners within an MSM network, we generated a synthetic population of 100,000 participants using a truncated Weibull distribution to represent the population of MSM and their number of sexual partners in a given geography. Similar methods have been described previously [\[6,](#page-15-5) [16\]](#page-17-4) to approximate the distribution of contacts. We then sampled this synthetic population with increasing sample sizes and calculated summary statistics. This process was repeated for 1,000 different synthetic populations.

 For each sample size, we calculated several summary statistics including the statistical coverage of the true mean number of sexual partners, the mean absolute error (MAE), and mean absolute percent error (MAPE). In order to assess the statistical coverage, the sam- pled values were fit using a Poisson model and the estimated rate parameter was recorded. The use of the Poisson distribution for contacts approximates an Erdos-Renyi model for the distribution of contacts in a random network [\[17\]](#page-18-0). Erdos-Renyi models are often used to parameterize network models for random network connections an have been used exten- sively in modeling sexual partnerships [\[13,](#page-17-1) [18\]](#page-18-1). To examine the results in the context of epidemiological estimates, we then used the properties of the sample to estimate the basic 103 reproduction number, R_0 , given a recover rate, γ , of 10 days and a contact rate, β , of 0.1%. Assuming that sexual partnerships are formed at random and in proportion to their count,

¹⁰⁵ we utilized the methodology described in Keeling and Rohani to estimate the contact rate $_{106}$ using the mean, M, and variance, V, of the number of contacts in the network as shown in ¹⁰⁷ equation [1](#page-6-1) [\[19\]](#page-18-2).

$$
R_0 = \frac{\mathbf{M}^2 + \mathbf{V}}{\mathbf{M}} \frac{\beta}{\gamma} \tag{1}
$$

¹⁰⁸ When generating synthetic populations, the heavy-tailed distribution was observed in the ¹⁰⁹ quantity of individuals who have a greater number of partners(Figure [1\)](#page-6-0).

Figure 1: Distribution of sexual partners drawn from a truncated Weibull distribution.

Results

 We found that poor coverage of the true mean number of partners was observed under all sampling strategies (Figure [2\)](#page-8-0).

 When examining the estimates, we find that surveys with smaller sample sizes consistently $_{114}$ underestimated the true number of sexual partners (Figure [5,](#page-11-0) panel A). Similarly, because of this underestimation of the mean number of partners as associated smaller estimated variance, the true basic reproduction number was generally underestimated (Figure [5,](#page-11-0) panel B). In essence, it is much more likely to sample individuals with below average number of sexual contacts than above average sexual contacts, which consistently biases estimates of the mean to be lower than the true value.

 When investigating the average expected bias of the estimate across all iterations of a given sample size, we found that there was no clear convergence to a single value for bias (Figure [4,](#page-10-0) panel A). Similarly when examining the average proportion of the true variance captured compared to the true variance, we find that the variance was consistently underestimated (Figure [4,](#page-10-0) panel B).

 However as the sample size increased both the mean absolute error and the mean absolute percent error began to converge after samples of 2,000 individuals were reached (Figure [3,](#page-9-0) $_{127}$ panels A and B).

Figure 2: Statistical coverage of the true mean number of sexual partners in the synthetic population based on survey sample size.

128 Discussion

 In our study of different surveying approaches of heavy-tailed networks, we found poor statistical coverage of the mean number of partners. We found that the average error, both on absolute (MAE) and proportional scales (MAPE) decreased rapidly after 2,000 individuals were sampled. Given the poor statistical coverage and consistent underestimation of the true variance of the network, estimates of the basic reproduction number for contact networks expected to have heavy tails are bound to be biased far below their true value. And the heavier the tail, the larger the bias in estimation of the basic reproductive number specifically and transmission rates more generally. In essence, the individuals most influential

Figure 3: The estimated and true values of the mean number of sexual partners, 95% confidence intervals from a Poisson regression shown (A). The calculated and true contact rate given synthetic epidemiological parameters (B).

¹³⁷ on transmission are scarce in the population and are unlikely to be sampled.

 These findings suggest that when surveying heavy-tailed social networks, many more individuals need to be sampled than is standard in most approaches assuming distributions closer to normal. As such, care must be taken to sample a sufficient number of the population in order to properly characterize these networks. These findings have strong implications for future field surveillance and modeling efforts.

Figure 4: The mean value of the bias for the average number of sexual partners based on the sample size (A) . The mean proportion of the variance in the number of sexual partners recovered as a function of the sample size (B).

Challenges of heavy-tailed distributions

 A key finding of this analysis is that because of extreme events, here sexual partners, the mean degree of the network will be poorly measured if insufficient samples of the population are taken. As demonstrated in our modeling study, this is further complicated by the fact that there did not appear to be convergence to a consistent value of bias as the sample size increase. Additionally, because of the large samples required to capture comparatively rare, extreme events, the variance estimates in smaller samples will fail to capture the true, larger variance of the population. As the proportion of the population in the heavy-tail of the distribution increases, the total variance will increase. Thus for the same estimator, more

Figure 5: The estimated and true values of the mean number of sexual partners, 95% confidence intervals from a Poisson regression shown (A). The calculated and true contact rate given synthetic epidemiological parameters (B).

 samples will be required in heavy-tailed distributions than distributions with thinner tails. Use of the Poisson distribution to model these effects will result in overconfidence in the uncertainty surrounding the estimates of the rate parameter due to the high variance of the true parameter despite growing sample sizes.

¹⁵⁶ Infectious disease dynamics

¹⁵⁷ Early studies of the dynamics of sexually transmitted infections in social networks resulted in ¹⁵⁸ discovery of reservoirs of infectious diseases due to heterogeneous mixing patterns of partners ¹⁵⁹ [\[20,](#page-18-3) [21\]](#page-18-4). Dense, heavy-tailed sexual networks allow for higher persistence of infection due

 to their scale free nature [\[22](#page-18-5)[–24\]](#page-19-0). This is in contrast to so called "small-world" networks which are best displayed as ring lattice where local connections tends to die out because of local pockets of immunity [\[16\]](#page-17-4). Additionally, sexual networks have been shown to be highly diffuse with partners traveling across domestic and international boundaries [\[25,](#page-19-1) [26\]](#page-19-2). Providing reliable estimates for these networks is vital in forming interventions to reduce the likelihood of outbreaks and controlling infection. Importantly, the role these networks play ₁₆₆ extends to understanding the dynamics of multi-drug resistance organisms such as Neisseria gonorrhoeae [\[27\]](#page-19-3) and sexually transmitted *Shigella flexneri* [\[28\]](#page-20-0). It is because of the long tails of these distributions that there is a higher potential risk for explosive outbreaks and persistence with the human population.

 Better characterization of these networks could be performed to supplement the informa- tion regarding the number of partners. Information about demographics, travel history, and other potential features of a given population could be used to supplement partner informa- tion to better capture the true network dynamics. For example, knowledge of participation in group sexual activities could allow modelers to better capture information about the long tail of sexual activities, estimate this population size and thus inform not only survey approaches but also better understand how disease may move within the network.

¹⁷⁷ Implications for modelers

 These findings suggest that higher rates of survey participants may be needed in order to more fully capture the dynamics of these heavy-tailed network. Because the infectious disease dynamics of these networks is largely driven by the heavy-tail of the distribution is it vital

 that care is taken to adequately capture the bias and variance. As we have shown in our simulation, depending on the sample size, the estimated basic reproduction number from a sample for a pathogen could be 5 times lower than the true value (Figure [5,](#page-11-0) panel A). This has large implications for the anticipated number of cases, potential number of hospitals, and depending on the virulence, the number of deaths. Furthermore, different control strategies might be adopted with information regarding the basic reproduction number [\[29\]](#page-20-1).

 Furthermore, modeling studies have shown that when risk-based models are used, the binning strategy for risk can have an impact on the inferences [\[30\]](#page-20-2). Furthermore, there are likely additional spatial, behavioral, and psycho-social phenomena that may need to be captured in order to capture how these social networks are connected. We did not assess the role of assortative mixing or homophily where those who are more like one another associate more strongly. Studies have shown that outbreak dynamics and effectiveness of different intervention strategies may differ depending on the overall contact structure and the degree of assortative mixing [\[31,](#page-20-3) [32\]](#page-21-0). As more data become available, these demographic, social, and spatial features could be user to better capture the tail behavior in models considering dense sexual networks. In the absence of data, modeler should be sensitive to the fact that existing data may categorically underestimate the true variance in networks and conduct sensitivity analyse to capture the fuller range of possibilities.

Financial support

 MD and CM receive support administered through the US Centers for Disease Control and Prevention's Epidemiology and Laboratory Capacity for the Prevention and Control of Infectious Diseases Cooperative Agreement (CK19-1904) for Strengthening the US Response to Resistant Gonorrhea.

Conflicts of interest

All authors have no relevant conflicts of interest to disclose.

Data availability

[A](https://github.com/wf-id/heavy-tail-sampling)ll data and code are stored in a publicly available GitHub repository ([https://github.](https://github.com/wf-id/heavy-tail-sampling)

```
208 com/wf-id/heavy-tail-sampling).
```
Author contribution

All authors contributed equally to the research design, data collection, formal analysis, in-

terpretation of the results, and writing of the paper.

²¹² References

16

²⁵⁰ Society, Interface 2, 295–307. issn: 1742-5689. pmid: <16849187> (Sept. 22, 2005).

18

- ²⁷⁴ 18. Datta, S., Mercer, C. H. & Keeling, M. J. Capturing sexual contact patterns in mod-
- ²⁷⁵ elling the spread of sexually transmitted infections: Evidence using Natsal-3. en. *PLoS*
- 276 ONE 13. Citation Key: dattaCapturingSexualContact2018. [https://www.ncbi.nlm.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6211691/)
- ²⁷⁷ [nih.gov/pmc/articles/PMC6211691/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6211691/) (2018).
- $_{278}$ 19. Keeling, M. J. & Rohani, P. *Modeling Infectious Diseases in Humans and Animals* ISBN:

²⁷⁹ 978-0-691-11617-4. [https://press.princeton.edu/books/hardcover/9780691116174](https://press.princeton.edu/books/hardcover/9780691116174/modeling-infectious-diseases-in-humans-and-animals)/ ²⁸⁰ [modeling-infectious-diseases-in-humans-and-animals](https://press.princeton.edu/books/hardcover/9780691116174/modeling-infectious-diseases-in-humans-and-animals) (2022) (Princeton Uni- $_{281}$ versity Press, Oct. 28, 2007).

- ²⁸² 20. Klovdahl, A. S. et al. Social Networks and Infectious Disease: The Colorado Springs 283 Study. Social Science & Medicine (1982) **38,** 79–88. ISSN: 0277-9536. pmid: <8146718> $_{284}$ (Jan. 1994).
- ²⁸⁵ 21. Stigum, H., Falck, W. & Magnus, P. The Core Group Revisited: The Effect of Partner ²⁸⁶ Mixing and Migration on the Spread of Gonorrhea, Chlamydia, and HIV. *Mathematical* 287 Biosciences 120, 1–23. ISSN: 0025-5564. [https://www.sciencedirect.com/science/](https://www.sciencedirect.com/science/article/pii/0025556494900361) ²⁸⁸ [article/pii/0025556494900361](https://www.sciencedirect.com/science/article/pii/0025556494900361) (2023) (Mar. 1, 1994).
- ²⁸⁹ 22. Pastor-Satorras, R. & Vespignani, A. Epidemic Spreading in Scale-Free Networks. ²⁹⁰ Physical Review Letters 86, 3200–3203. [https : / / link . aps . org / doi / 10 . 1103 /](https://link.aps.org/doi/10.1103/PhysRevLett.86.3200) ²⁹¹ [PhysRevLett.86.3200](https://link.aps.org/doi/10.1103/PhysRevLett.86.3200) (2023) (Apr. 2, 2001).

20

- 28. Thorley, K. *et al.* Emergence of extensively drug-resistant and multidrug-resistant Shigella flexneri serotype 2a associated with sexual transmission among gay, bisex- ual, and other men who have sex with men, in England: a descriptive epidemiological ³¹⁹ study. en. The Lancet Infectious Diseases 23. Citation Key: thorleyEmergenceExten- sivelyDrugresistant2023a, 732–739. issn: 1473-3099. [https://www.sciencedirect.](https://www.sciencedirect.com/science/article/pii/S1473309922008076) [com/science/article/pii/S1473309922008076](https://www.sciencedirect.com/science/article/pii/S1473309922008076) (June 2023).
- 29. Fraser, C., Riley, S., Anderson, R. M. & Ferguson, N. M. Factors That Make an Infec- tious Disease Outbreak Controllable. Proceedings of the National Academy of Sciences 101, 6146–6151. issn: 0027-8424, 1091-6490. pmid: <15071187>. [https://www.pnas.](https://www.pnas.org/content/101/16/6146) [org/content/101/16/6146](https://www.pnas.org/content/101/16/6146) (2020) (Apr. 20, 2004).
- 30. Suen, S.-c., Goldhaber-Fiebert, J. D. & Brandeau, M. L. Risk Stratification in Com-
- ³²⁷ partmental Epidemic Models: Where to Draw the Line? *Journal of theoretical biology* 428, 1–17. issn: 0022-5193. pmid: <28606751>. [https://www.ncbi.nlm.nih.gov/pmc/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5555315/) [articles/PMC5555315/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5555315/) (2023) (Sept. 7, 2017).
- 31. Gupta, S., Anderson, R. M. & May, R. M. Networks of sexual contacts: implications δ ₃₃₁ for the pattern of spread of HIV. eng. *AIDS (London, England)* 3. Citation Key: gup-
- taNetworksSexualContacts1989, 807–817. issn: 0269-9370 (Dec. 1989).

