Appendix B

This appendix offers a supplementary test of a sampling approach. In the main analysis, we focus on a single sampling scenario: where a researcher samples a large portion of a small network. We take a 25% sample from a network of size 700. Here, we expand the analysis, replicating the results with a much larger network and a much lower sampling rate. In this way, we can assess the validity of a sampling approach under more difficult conditions.

The network of interest comes from Project 90, the Colorado Springs study of high-risk individuals (see Morris and Rothenberg 2011 for the data source). The population of interest includes at-risk individuals for HIV transmission, including drug injectors and sex workers. Researchers attempted to saturate the population in this city and we treat the network as a full census. The data include social connections based on sex, needle sharing and social ties. The true network includes 5492 nodes and 21644 edges. We base our analysis on a 5% sample of the network.

The test presented here is a difficult one, as we use a larger network and a lower sampling rate than in the main analysis. Additionally, the properties of the network make this a particularly difficult test of a sampling approach. First, the network has high transitivity (.37) and high average degree (7.88), and we have already seen that the bias is higher in such networks. Second, the network has a skewed degree distribution. Inference is harder when the degree distribution is skewed: a few actors have disproportionately high degree, yet they are no more likely to be sampled than any other node. High degree nodes thus have a large impact on network structure, but are often missed in a random sample (see Smith 2015). And third, the network is disconnected, with 20% of the nodes outside the main component (a component is a set of nodes connected by at least one path; the main component is the largest set of nodes connected by at

least one path). The diffusion simulation will be highly variable under such conditions. Global diffusion is possible (albeit not necessary) when the initial seed is in the main component; in contrast, global diffusion is impossible if the initial seed is not in the main component (as they are disconnected from the rest of the network and cannot pass the product beyond their own borders). The results are thus highly dependent on the initial seed, making inference more difficult.

The analysis is the same as before. The agent-based model of diffusion follows a simple contagion process, with three adoption probabilities: .1, .2, and .3. We again take 100 independent samples. We assume that the data collected have the same pattern as in the main text. The only difference is in the demographic characteristics assumed to be collected. Here, the characteristics of interest include: race, gender, employment status, and illicit activity (drug dealer, sex worker, pimp or none).

We present the results below in Figure B1. The results follow the same form as in Figures 8-11. There are three subplots, one for each adoption probability. Each subplot shows the true proportion adopting, the mean estimate and the error bounds. The estimate are, in general, quite good, despite the difficulty of the test. Looking at the high adoption results, the median bias over the 30 time periods is under 2%. The results are on par with the findings in the main text. For example, for period 20, the mean estimate is .424, while the true value is .421 (a relative bias under 1%). The estimates are, as expected, more uncertain than before. The median standard error (over the 30 time periods) is .06, higher than with any network used in the main analysis. For time period 20, 95% of the estimates fall between .34 and .56, a wide range of values. We see similar results with the lower adoption probabilities, although the bias is higher here. The median relative bias (over the 30 time periods) is about .037 for the medium adoption analysis.

For example, for period 20, the true proportion adopting is .368, the mean estimate is .374, and 95% of the estimates fall between .289 and .488.

Overall, the results suggest that it is possible to produce good approximations of the true diffusion curves using a small sample on a large network. The caveat is that the estimates can be quite uncertain, with high variability sample-to-sample. A researcher concerned with the variability of the estimates would have to sample more than 5% of the network.