Web Appendix Content

Social network clustering of sexual violence experienced by adolescent girls

Web Appendix 1: Description of sampling procedure and response rates

Web Appendix 2: Description of procedure for generating dyadic data

Web Appendix 3: Measures

Web Appendix 4: Detailed on the use of item response theory to develop a risk score, and calculation of eigenvector centrality.

Web Table 1: Item Response Theory Parameters

Web Table 2: Dyadic level bivariate linear regression analyses to assess association between ego's and alter's risk scores

Web Table 3: Dyadic level bivariate logistic regression analyses to assess social network level risk factors on a girls probability of experiencing sexual violence by Wave II

Web Table 4: Dyadic level bivariate analyses using logistic regression to assess social network level risk factors on a girls probability of experiencing sexual violence by wave 4

Web Table 5: Logistic regression models showing the association between sexual violence (WI/II) and network centrality WII above mean, bivariate models

Web Appendix 1

Sampling procedure and response rates

The core in-home sample is comprised of 12,105 respondents. These respondents were randomly selected from a stratified sampling design based on school, grade, and sex, which makes this sample nationally representative and "self-weighting." An additional 8,000 or so respondents were sampled as part of special supplemental groups. These oversampled groups include Chinese, Cuban, Puerto Rican, high educated Black, and disabled students, siblings, and a saturated sample in which every student in 16 different schools were interviewed. About 79% of those invited to become part of the in-home sample did, yielding a total sample size of 20,745. Sampling weights were created to adjust for the oversampling of these populations within the in-home sample. For more detail on the sampling procedure and response rates, see Harris (2013) [1]

Web Appendix 2

Procedure for generating dyadic data

1. Transform the Wave I *In-Home* sample of adolescents and the Wave II *In-Home* sample of adolescents into a single sample of adolescent-friend directed dyads based on a question asking respondents to list up to five female and five male friends.
    a. Merge the dyadic data generated from the Wave I *In-Home* sample and the Wave II I*n-Home* sample to generate a comprehensive list of unique friendship dyads reported in Wave I and Wave II. Some friendship dyads will only be reported in Wave I, some will only be reported in Wave II, and some will be reported in both. These latter dyads are combined in the merged data so that they are not double counted.
    b. A pair of friends could be listed twice if both respondents report the other person as a friend (i.e., "friend 1-friend 2" dyad AND "friend 2-friend 1" dyad)
    c. This transformation of the data increases the sample size (although not by 5-fold since not every female respondent lists 5 female friends); the size of the sample is later reduced by eliminating dyads that do not meet the study's criteria (see step 3 below).
2. Information is added to the dyadic data using the adolescents' and the friends' responses to both the *In-Home* questionnaire. Specifically:
    a. Network measures (e.g., centrality) are created based on the full network information of each school that was generated from respondents' listing of 10 friends (from the *In-Homel* questionnaire (for Waves I and II).
    b. Measures of the experience of sexual violence are added for both the adolescent and the friend based on the *In-Home* questionnaire (for Waves 1, Waves 2, and Wave 4).
    c. Measures of substance use are created for both the adolescent and the friend based on the *In-Home* questionnaire (for Waves 1, Waves 2, and Wave 4).
3. The size of the dyadic data is reduced by dropping observations for the following reasons:
    a. Either the adolescent or the friend is not part of the *In-Home* sample.
    b. The adolescent-friend pairing is male-male or female-male or male-female.
    c. The adolescent or the friend did not answer a question about sexual violence, substance use, or one of the control variables (i.e., listwise delete of observations due to missingness from item non-response).
4. The final sample is 3139 girl dyads based on 1658 individual interviews

Web Appendix 3

Measures

Adolescent history of alcohol use was assessed in Waves I and II by asking "Over the past 12 months, on how many days did you drink alcohol?" Binge drinking was measured by asking "Over the past 12 months, on how many days did you drink five or more drinks in a row?" Respondents were also asked whether "you did something you regretted because you had been drinking" and whether "you got into a sexual situation that you later regretted because you had been drinking." Tobacco use was measured by asking "During the past 30 days, on how many days did you smoke cigarettes?" Respondent marijuana use was measured by asking "During the past 30 days, how many times have you used marijuana?".

In Wave IV alcohol use was measured by asking respondents "Have you ever continued to drink after you realized drinking was causing you any emotional problems (such as feeling irritable, depressed, or uninterested in things or having strange ideas) or causing you any health problems (such as ulcers, numbness in your hands/feet or memory problems)?" Marijuana use was measured using a similar question that asked "Have you ever continued to use marijuana after you realized using marijuana was causing you any emotional problems (such as feeling depressed or empty, feeling irritable or aggressive, feeling paranoid or confused, feeling anxious or tense, being jumpy or easily startled) or causing you any health problems (such as persistent cough, sore throat or sinus problems, heart pounding, headaches or dizziness, or sexual difficulties)?"

Web Appendix 4

Item response theory

An item response theory (IRT) model is a latent variable model that is similar to a factor analysis model. The latent variable model estimates the "true" level of some unobservable trait using a set of binary indictors that are related to that trait. The use of these types of models is now quite common in the social and psychological sciences because they offer a principled method for uncovering latent characters of the units within a population such as general intelligence[2,3], health outcomes[4,5], and many any other unobservable or difficult to observe behaviors and traits[6,7]. The latent variable model we use incorporates the six binary survey response variables and quantifies the uncertainty of each subject estimate, conditional on the availability of each variable included in the model. Thus, a useful feature of the Bayesian version of the model we use is that missing data does not lead to a loss of observations but only increases the uncertainty for the estimate for that subject.

Eigenvector centrality

While this centrality is an intuitive way to think about which people might be better connected, it yields a practical problem — how do we simultaneously estimate the centrality of a given respondent and the centralities of the friends to whom she is tied? Let x be a vector of centrality scores so that each person's centrality x_j is the sum of the centralities of her friends: $x_j = a_{1j}x_1 + a_{2j}x_2 + \cdots + a_{nj}x_n$. This yields *n* equations that we can represent in matrix form as $x = A^T x$. It is unlikely that these equations have a nonzero solution, so Bonacich[8] suggests an important modification. Suppose the centrality of a respondent is proportional to instead of equal to the centrality of her friends. Then $\lambda x_i = a_{1i}x_1 + a_{2i}x_2 + \cdots + a_{ni}x_n$ which can be represented as $\lambda x = A^T x$. The vector of centralities x can now be computed, since it is an eigenvector of the eigenvalue λ. Although there are *n* nonzero solutions to this set of equations, in practice the eigenvector corresponding to the principal eigenvalue is used because it maximizes the accuracy with which the associated eigenvector can reproduce the adjacency matrix[9].

Web Table 1: Item Response Theory Parameters

	Alpha	Alpha sd	Beta	Beta sd
Alcohol Use	-1.008	(0.112)	10.828	(0.767)
Tobacco Use	-1.230	(0.019)	1.279	(0.020)
Marijuana Use	-2.336	(0.027)	1.369	(0.024)
Binge Drinking	-3.997	(0.130)	5.678	(0.181)
Regret because of Drinking	-3.774	(0.062)	2.586	(0.054)
Regret Sexual Behavior	-4.806	(0.088)	2.565	(0.064)

Web Table 1 displays alpha and beta parameters from six logistic regression equations (model parameters), which each link the observed risk behavior with the underlying estimate of the latent risk score for each subject. All of the parameters, the latent variable, and the model parameters are estimated using Bayesian simulation. The negative alpha parameters indicate low base line probabilities for engaging in each behavior for the sample of subjects in the model. The large beta parameters indicate that each behavior contributes important information to the estimation of the underlying latent characteristic.

Web Table 2: Dyadic level bivariate linear regression analyses to assess association between ego's and alter's risk scores*

	Ego's risk score WVII		
	Beta	SE	P
Risk score alter WVII	0.39	0.02	0.00

*multiple observations per individual controlled for using general estimating equations

Web Table 3: Dyadic level bivariate logistic regression analyses to assess social network level risk factors on a girls probability of experiencing sexual violence by Wave II*

	Beta	SE	P	Beta	SE	P	Beta	SE	P
Alter Sexual Violence by WVII	0.84	0.22	0.00						
Risk score ego WVII				0.56	0.08	0.00			
Risk score alter WVII							0.34	0.06	0.00

*multiple observations per individual controlled for using general estimating equations

Web Table 4: Dyadic level bivariate analyses using logistic regression to assess social network level risk factors on a girls probability of experiencing sexual violence by wave 4*

	Beta	SE	P
Alter Sexual Violence by WVII	0.41	0.21	0.06
Risk score ego WVII	0.12	0.08	0.16
Risk score alter WVII	-0.02	0.06	0.69
Problem Marijuana 4	0.77	0.37	0.04
Problem Drinking 4	0.92	0.34	0.01
Smoking 4	0.55	0.18	0.00

*multiple observations per individual controlled for using general estimating equations

Web Table 5 Individual logistic regression models showing the association between sexual violence (WI/II) and network centrality WII above mean, bivariate models

	Total degree	In-degree	Out-degree	Eigenvector
Sexual Violence				
Beta	-0.39	-0.38	-0.38	-1.14
SE	0.24	0.24	0.24	0.40
P	0.10	0.10	0.12	0.00

References

1. Harris K. The Add Health Study:
Design and Accomplishments. Chapel Hill, NC, 2009.
2. Rasch G. Probabilistic models for some intelligence and attainment tests: ERIC; 1993.
3. Lord FM. Applications of item response theory to practical testing problems: Routledge; 1980.
4. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research* 2007; **16**(1): 5-18.
5. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological medicine* 2002; **32**(06): 959-76.
6. Schnakenberg KE, Fariss CJ. Dynamic patterns of human rights practices. *Political Science Research and Methods* 2014; **2**(01): 1-31.
7. Borsboom D. Measuring the mind: Conceptual issues in contemporary psychometrics: Cambridge University Press; 2005.
8. Bonacich P. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology* 1972; **2**(1): 113-20.
9. Bonacich P. Power and centrality: A family of measures. *American journal of sociology* 1987: 1170-82.