Supporting Information

Fowler *et al.* 10.1073/pnas.0806746106

SI Text

The Add Health Data. The National Longitudinal Study of Adolescent Health (1) (Add Health) is a study that, among other topics, explores the causes of health-related behavior of adolescents in grades 7 through 12 and their outcomes in young adulthood. Three waves of the Add Health study have been completed: Wave I was conducted in 1994–1995, Wave II in 1996, and Wave III in 2001–2002.

In Wave I of the Add Health study, researchers collected an “in-school” sample of 90,118 adolescents chosen from a nationally-representative sample of 142 schools. These students filled out questionnaires about their friends, choosing up to 5 male and 5 female friends who were later identified from school-wide rosters to generate information about each school’s complete social network. We use these nominations to measure the in-degree (the number of times an individual is named as a friend by other individuals) and out-degree (the number of individuals each person names as a friend) of each subject. The in-degree is virtually unrestricted (the theoretical maximum is $N - 1$, the total number of other people in the network) but the out-degree is restricted to a maximum of 10 due to the name generator used by Add Health. Fortunately, most subjects (90.0%) named fewer than the maximum, and there is substantial variation in the total number of friends named by each person (mean = 3.8, standard deviation = 3.7).

We also measure transitivity as the empirical probability that any one of an individual’s friends names or is named by a friend by any of that individual’s other friends. This is just the total number of triangles of ties divided by the total possible number of triangles for each individual.

Finally, we measure betweenness centrality which identifies the extent to which an individual in the network is critical for passing support or information from one individual to another. If we let σ_{ik} represent the number of shortest paths from subject i to subject k, and σ_{ijk} represent the number of shortest paths from subject i to subject k that pass through subject j, then the betweenness measure x for subject j is

$$x_j = \sum_{i \neq j \neq k} \frac{\sigma_{ijk}}{\sigma_{ik}}.$$

Note that for the purpose of measuring transitivity and betweenness centrality, we assume all directed ties are undirected, so that a tie in either direction becomes a mutual tie. For example, we consider the case where A names B, B names C, and C names A to be transitive. Likewise, if A names B, A names C, and B names C, we consider the relationships to be transitive for all three individuals.

The Add Health team created a genetically informative sample of sibling pairs, including all adolescents that were identified as twin pairs, half-siblings, or unrelated siblings raised together. Twins and half biological siblings were sampled with certainty. The Wave I sibling-pairs sample has been found to be similar in demographic composition to the full Add Health sample (2).

The Twin Study Method. To estimate the heritability of egocentric social network attributes, we study the patterns of same-sex (identical) monozygotic (MZ) twins who were conceived from a single fertilized egg and same-sex (non-identical) dizygotic (DZ) twins who were conceived from two separate eggs. MZ twins share 100% of their segregating genes, while DZ twins share only 50% on average. Thus, if network attributes are heritable, MZ twins should exhibit more similarity than DZ twins. Moreover, if it is assumed that MZ twins and DZ twins share comparable environments, then we can use these concordances to estimate explicitly the proportion of the overall variance attributed to genetic, shared environmental, and unshared environmental factors. Very few differences have been found between twins and non-twins, therefore we expect the results for twins to be generalizable to a non-twin population (3).

Some scholars have objected to the assumption that MZ and DZ environments are comparable, arguing that the identical nature of MZ twins cause them to be more strongly affiliated and more influenced by one another than their non-identical DZ counterparts. If so, then greater concordance in MZ twins might merely reflect the fact that their shared environments cause them to become more similar than DZ twins. However, studies of twins raised together have been validated by studies of twins reared apart (4), suggesting that the shared environment does not exert enhanced influence on MZ twins. More recently, Visscher *et al.* use the small variance in percentage of shared genes among DZ twins to estimate heritability without using any MZ twins, and they are able to replicate findings from studies of MZ and DZ twins reared together (5). Moreover, personality and cognitive differences between MZ and DZ twins persist even among twins whose zygosity has been miscategorized by their parents (6), indicating that being mistakenly treated as an identical twin by one’s parents is not sufficient to generate the difference in concordance. And, although MZ twins are sometimes in more frequent contact with each other than DZ twins, it appears that twin similarity (e.g., in attitudes and personality) may cause greater contact rather than vice versa (7). Finally, contrary to the expectation that the influence of the unshared environment would tend to decrease concordance over time, once twins reach adulthood, MZ twins living apart tend to become more similar with age (6).

The Add Health data has been used in a wide variety of twin studies (8). As a result, there have been several analyses of the comparable environments assumption for MZ and DZ twins. One of these studies reported that the environments were not comparable (9), but other scholars have pointed to serious deficiencies in this work that negated its conclusions (10). For example, Horwitz *et al.* (9) showed that including observed social variables in a twin model causes the P value on the genetic component for males trying alcohol to change from being just below 0.05 to just above it. Freese and Powell (10) note that this is unsurprising since adding variables to a regression can have a substantial effect on efficiency. Even worse, they point out that Horwitz *et al.* (9) do not acknowledge that their own fit statistics indicate the models with and without social variables are statistically indistinguishable, suggesting that the model with additional variables should be rejected.

The twin study design has been used frequently to identify the relative degree to which genetic and environmental factors influence an observed outcome (11–12). The basic twin model assumes that the variance in observed behavior can be partitioned into additive genetic factors (A), and environmental factors which are shared or common to co-twins (C), and unshared environmental (E). This is the so-called ACE model. The role of genotype and environment are not measured directly but their influence is inferred through their effects on the covariances between twin siblings (12). No observed covariates are needed in the model because the degree to which they contribute to variance is a part of one of three variance components (A, C, and E). More formally, these components are derived from known relationships between three observed statistics (11):

$$\sigma_{\mathrm{P}}^2 = \sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{C}}^2 + \sigma_{\mathrm{E}}^2$$

$$\mathrm{COV}_{\mathrm{MZ}} = \sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{C}}^2$$

$$\mathrm{COV}_{\mathrm{DZ}} = 0.5\sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{C}}^2$$

In this equation, σ_{P}^2 is the observed phenotypic variance (the same for monozygotic and dizygotic twins); $\mathrm{COV}_{\mathrm{MZ}}$ and $\mathrm{COV}_{\mathrm{DZ}}$ are the observed covariances between monozygotic twins who share all their genes and dizygotic twins who share only half on average; and σ_{A}^2, σ_{C}^2, and σ_{E}^2 are the variance components for genes, common environment, and unshared environment, respectively. This is a system of three equations and three unknowns so it is identified:

$$\begin{pmatrix} \sigma_{\mathrm{A}}^2 \\ \sigma_{\mathrm{C}}^2 \\ \sigma_{\mathrm{E}}^2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 0.5 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{COV}_{\mathrm{MZ}} \\ \mathrm{COV}_{\mathrm{DZ}} \\ \sigma_{\mathrm{P}}^2 \end{pmatrix}$$

Heritability, or the proportion of the variance explained by genetic factors, can be estimated as $\sigma_{\mathrm{A}}^2/(\sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{C}}^2 + \sigma_{\mathrm{E}}^2)$. We use the software package MX to estimate this structural equations model (13). Table S1 shows the complete results for each of the social network measures of interest.

Since the variance components are not directly observable, the ACE model's assumption of additivity cannot be tested and more complicated relationships are possible. For example, it is possible that genes interact with the environment (GxE) or with other genes (GxG) to yield variation in behavior, or at a higher level, phenotypes interact with the environment (PxE) (13). We limit our analysis to the ACE model but point out that if a strong effect for genes is found in the additive model, then genes are also likely to play a role in more complex specifications as well.

Finally, it is important to clarify the difference between the common environment (C) and the unshared environment (E) in the twin model. Common environment includes the family environment in which both twins were raised, as well as any other factor to which both twins were equally exposed. In contrast, the unshared environment includes idiosyncratic influences that are experienced individually. It is possible to have unshared environmental exposure as a child (e.g., twins may have different friends with different beliefs) and to have shared environments as an adult (e.g., twins may share the same friend). Thus, the distinction between common and unshared environment does not correspond directly to family-nonfamily or adult-child differences in factors that influence a given behavior. Moreover, there may be a similarity in the objective environment but twins may have idiosyncratic experiences that influence their effective environment, and these idiosyncratic experiences may create an unshared rather than a common environmental influence on variation in the phenotype (13).

Methods for Generating Networks from Extant Models. ***Erdos–Renyi.*** We assume there are N nodes and E edges. The probability of a social tie from i to j is $E/(N(N - 1))$ (14).

Fitness. We assume there are N nodes and E edges. A node i is drawn at random and added at each integer time t, each with a fitness $h_{\mathrm{i}} \sim \mathrm{Uniform}[0,1]$. The probability the new node i attaches to any existing node j is

$$(E/N)\left(d_j h_j / \sum_{m=1}^{t} d_m h_m\right),$$

where d_j is the degree of node j at the time i is added to the network (15). Note we do not treat t in this model as an intrinsic characteristic. This means that when we implement the mirror network method, we do not force each twin in a pair to enter the network at the same time t.

Social Space. We assume there are N nodes and E edges. Each node is placed in a one dimensional social space on the unit interval with uniformly distributed probability. The locations h_i of node i and h_j of node j determine the probability of a mutual social tie between them that equals $1/(1 + (|h_i - h_j|/\beta)^\alpha)$. The parameters $\alpha = 1.45$ and $\beta = 0.00115$ were chosen to generate the empirically observed mean degree (E/N) and transitivity (16).

ERGM. we assume there are N nodes and E edges. We assume a node characteristic $h_{\mathrm{i}} \sim \mathrm{Uniform}[0,1]$ is correlated with in-degree, and we assume the number of transitive triplets is the number observed in the network. Using the ergm function in the STATNET package (17), we constrain the model to have the observed number of nodes and edges and we choose coefficients for the in-degree characteristic and for the generation of transitive triplets that matches the observed transitivity and heritability of in-degree as closely as possible to the observed transitivity and heritability in the Add Health networks (in our optimization, we find that a coefficient of 2.5 for the in-degree and 2.7 for the transitive triplets performs best). We then simulate networks from this model, keeping the distribution of node characteristics fixed across all simulated networks (18).

For additional details on how these models were implemented and specific results, see *SI Appendix*.

Fingerprint Procedure for Assessing the Probability of a Given Distribution of 3-Motifs and 4-Motifs. The structure of social networks can be characterized by the distribution of k-motifs, or isomorphic combinations of ties between all sets of k nodes (19–22). In a directed network, there are 16 possible combinations of social ties among 3 nodes and 218 possible combinations of social ties among 4 nodes. Several procedures have been proposed for identifying significant motifs (19–22), but our goal here is to use the motifs to determine which proposed model is most likely to generate a network like the one observed. To do this, we simulate 100 networks from each proposed model using the empirical distribution of nodes and edges in each of the largest 100 Add Health networks (we restrict attention to the largest 100 networks to minimize noise that results from an inadequate number of observations in the smaller networks). We then count the total number of motifs of each type in each network and divide by the total number of motifs in that network to generate the empirical probability that k nodes form any given motif (the motif "fingerprint" of the network). For each motif, we calculate the mean (μ) and variance (σ^2) of the motif probability across all 100 simulated networks. We then use the mean and variance to estimate the α and β parameters of one dimension of a multivariate beta density that characterizes the distribution of motif probabilities: $\alpha = \mu^2(1 - \mu)/\sigma^2 - \mu$; $\beta = \mu(1 - \mu)^2/\sigma^2 - (1 - \mu)$. The likelihood of observing a given motif probability can then be estimated from the value of a beta distribution with parameters α and β at the point of the observed motif probability. The likelihood of observing a full set of motifs is the product of the likelihoods for each possible motif.

It is important to note that we make a strong assumption with this method that the observed motif probabilities are independent of one another. However, the procedure shows excellent discriminatory power despite this strong assumption. We simulated 100 networks from each of the proposed models and used the fingerprint likelihood method we have described to test whether the procedure assigned the highest likelihood to the model that generated the simulated network. Fig. S1 shows the results. The label at the left indicates which model was used to generate the simulated networks, Fig. S1 *Right* shows the likelihoods for the 3-motif fingerprint and shows the likelihoods for the 4-motif fingerprint. For ease of exposition, we show adjusted likelihoods $-\log(c - LL)$, where c is a constant across all networks and models and LL is the log likelihood of generating an observed network. Each point in the figure represents the adjusted likelihood that a proposed model generated the simulated network.

In all 100 cases for each observed network and for each motif structure, the model with the highest likelihood was the one that generated the data. We also show the likelihood generated by the set of 100 observed networks as well for comparison. We repeated this exercise 10 times (only 1 repetition shown), generating 10 repetitions × 2 fingerprints × 100 networks × 5 models = 10,000 tests of the procedure, and successfully identified the model that generated the network model in all 10,000 cases.

1. Udry JR (2003) *The National Longitudinal Study of Adolescent Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002 [machine-readable data file and documentation].* (Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC).
2. Jacobson KC, DC Rowe (1998) Genetic and Shared Environment Influences on Adolescent BMI: Interaction with Race and Sex. *Behav Gen* 28:265–275.
3. Kendler KS, Nick G, Martin ACH, Eaves LJ (1995) ''Self-report psychiatric symptoms in twins and their nontwin relatives: Are twins different? *Am J Med Gen Neuropsychiatric Gen* 60:588–591.
4. Bouchard TJ (1998) Genetic and environmental influences on adult intelligence and special mental abilities. *Hum Biol* 70:257–279.
5. Visscher PM, *et al.* (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* 2:e41.
6. Bouchard TJ, McGue M (2003) Genetic and environmental influences on human psychological differences. *J Neurobiol* 54:4–45.
7. Posner S, Baker LA, Martin NG. (1996) Social contact and attitude similarity in australian twins. *Behav Gen* 26:123–134.
8. Harris KM, Halpern CT, Smolen A, Haberstick BC (2006) The national longitudinal study of adolescent health (Add Health) twin data. *Twin Res Hum Gen* 9:988–997.
9. Horwitz AV, Videon TM, Schmitz MF (2003) Rethinking twins and environments: Possible social sources for assumed genetic influences in twin research. *J Health Soc Behav* 44:111–129.
10. Freese J, Powell B (2003) Tilting at twindmills: Rethinking sociological responses to behavioral genetics. *J Health Soc Behav* 44:130–135.
11. Evans DM, Gillespie NA, Martin NG (2002) Biometrical genetics. *Biol Psychol* 61:33.
12. Neale MC, Cardon LR (1992) *Methodology for Genetic Studies of Twins and Families.* (Kluwer, Dordrecht, The Netherlands).
13. Bouchard TJ, Lykken DT, McGue M, Segal NL, Tellegen A (1990) Sources of human psychological differences: the Minnesota Study of Twins Reared Apart. *Science* 250:223.
14. Erdős P, Rényi A (1959) On Random Graphs. I. *Publicationes Mathematicae* 6:290–297.
15. Bianconi G, Barabási AL (2001) Competition and multiscaling in evolving networks. *Europhys Lett* 54: 436.
16. Boguñá M, Pastor-Satorras R, Díaz-Guilera A, Arenas A (2004) Models of social networks based on social distance attachment. *Phys Rev E* 70:056122.
17. Handcock MS, Hunter DR, Butts CT, Goodreau SM, Morris M (2003) ergm: A Package To Fit, Simulate and Diagnose Exponential-Family Models for Networks, Version 2 (Statnet Project, Seattle, WA).
18. Snijders TAB, Pattison PE, Robins GL, Handcock MS (2006) New specifications for exponential random graph models. *Sociol Methodol* 36:99–153.
19. Middendorf M, Ziv E, Wiggins CH (2005) Inferring network mechanisms. *Proc Natl Acad Sci USA* 102:3192–3197.
20. Milo R, *et al.* (2002) Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298:824–827.
21. Holland P, Leinhardt S (1975) in *Sociological Methodology*, ed Heise D (Jossey-Bass, San Francisco), pp 1–45.
22. Wasserman S, Faust K (1994) *Social Network Analysis* (Cambridge Univ Press, New York).

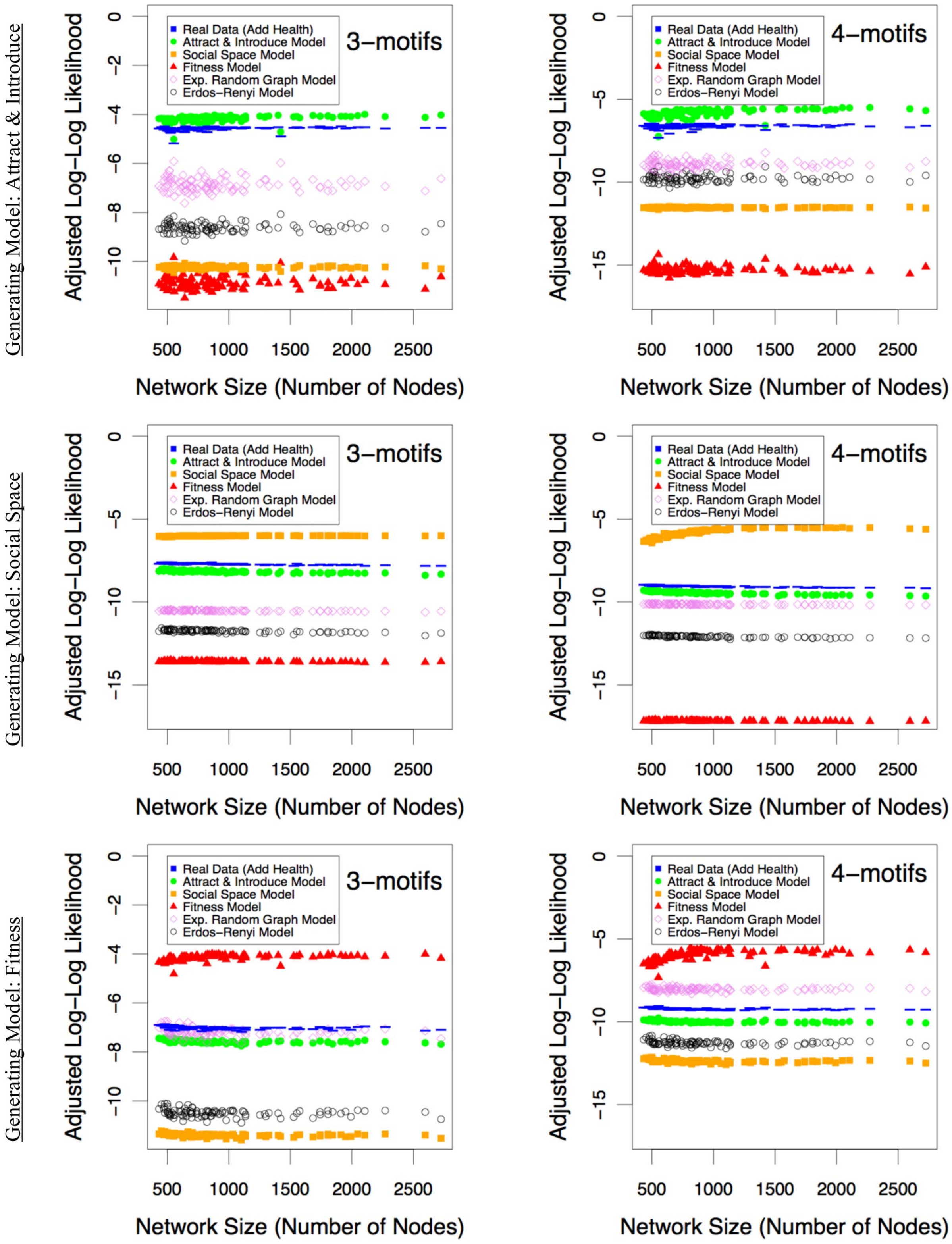

Fig. S1. Fit of each model to simulated networks.

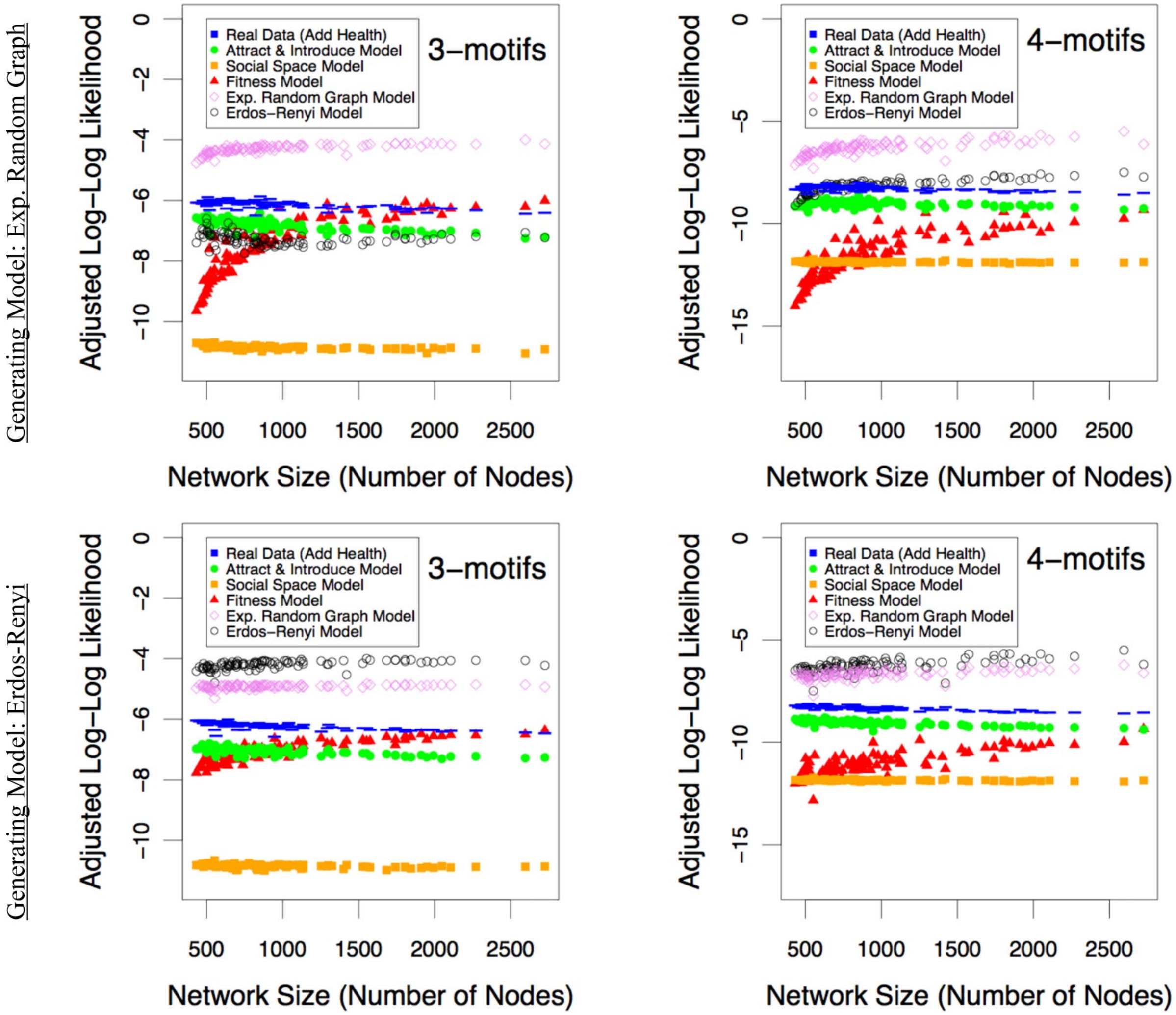

Fig. S1 (continued).

Table S1. ACE model estimates from the Add health data

	Proportion of Variance Explained by			
Model	Genetic Factors	Common Environment	Unshared Environment	Model fit (−2LL)
In-degree	0.46 (0.23, 0.69)	0.21 (0.00, 0.40)	0.34 (0.28, 0.40)	2,386.11
Transitivity	0.47 (0.13, 0.65)	0.09 (0.00, 0.36)	0.44 (0.35, 0.56)	2,033.91
Betweenness centrality	0.29 (0.05, 0.39)	0.00 (0.00, 0.19)	0.71 (0.61, 0.81)	2,489.30
Out-degree	0.22 (0.00, 0.47)	0.16 (0.00, 0.40)	0.63 (0.53, 0.75)	2,284.09
Out-degree, dropping those who name 10 friends	0.00 (0.00, 0.39)	0.44 (0.09, 0.52)	0.56 (0.46, 0.67)	1,996.58

Note. 95% confidence intervals indicated in parentheses beneath each estimate. First four models based on 307 monozygotic and 248 dizygotic same-sex twin pairs. The last model drops subjects who named the maximum 10 possible friends, yielding 256 monozygotic and 204 dizygotic same-sex twin pairs. Network measures were transformed to have zero mean and unit variance within each school network to prevent differences between schools from influencing the results.

Other Supporting Information Files

SI Appendix