SOCIUS
Original Article

AMERICAN SOCIOLOGICAL ASSOCIATION

Socius: Sociological Research for a Dynamic World
Volume 5: 1–7

DOI: 10.1177/2378023119845758
srd.sagepub.com
SAGE

Local Network Regressions

Peter McMahan[1], Adam Slez[2], and John Levi Martin[3]

Abstract

The authors propose and illustrate an exploratory technique to shed light on the degree to which bivariate relations between individual-level variables themselves vary over a network. The authors discuss limitations and possible extensions.

Keywords

network regression, local regression, variance

Network researchers, like other social scientists, are often interested in the covariation between measured variables, though in their case, these variables make reference to a social structure that can be interpreted as a graph. In most cases, they are particularly interested in variables that are themselves relational, such as the existence or quality of some tie. However, they may also be interested in the covariation of individual-level variables but attentive to the location of persons in a social network.

One way in which this is generally done is to take network-based attributes of nodes and treat them as individual-level variables. For example, we might take graph-theoretic quantities such as eigenvector centrality, or contextual variates such as the proportion of those to whom a node is tied that are in some measured state, and enter these in individual-level regressions. However, another possible approach would be to explore heterogeneity in a pattern of covariation at different parts of the network.

For example, consider a network within a large high school. We might, in examining this high school, be interested in whether school involvement (e.g., the number of clubs a student belongs to) predicts academic achievement or whether popularity is positively associated with grade point average (GPA). We might of course compare the coefficients from one high school with those from another. But we might also be interested in making internal comparisons with different sets of students within the school. In some cases, we have a priori theoretical knowledge of or interest in some categorical variables that might predict the strength of this relation (say, that between academic achievement and extracurricular activity). For example, we might wonder whether this effect was stronger among boys or among girls. In such cases, we might incorporate the categorical predictor as an interaction in an equation predicting achievement on extracurricular activities. However, in other cases, we may suspect that such categories, rather than being unmoved movers, have effects that themselves vary across the social structure of the school. We might, then, suspect that it is location in the friendship (say) network of students that predicts the nature of the relation between two variables: perhaps in some circles, extracurricular activities are positively associated with achievement and, in other circles, negatively. Similarly, perhaps in some circles, girls tend to be higher achievers than boys, whereas in other circles, the reverse is true.

If we were to divide the school up into exclusive and exhaustive subsets of friends, we might treat membership in some group such as race or sex or some other categorical variable. However, in most cases, this requires making "cuts" in a larger graph whereby we decide to treat some relations as if they did not exist, simply because if they did not, we could treat these subsets as separable. But this may be to throw away information that is crucial to reproducing how each student perceives his or her local environment. We might be better off leaving the network as observed and attempting to describe the full range of association between our variables at all positions of the network. We then can see whether, from the perspective of each individual, it would appear that the variables are related and, if so, to what degree. It is for this reason that we here propose a technique of carrying out what we call local network regressions, in effect

[1]McGill University, Montreal, QC, Canada
[2]University of Virginia, Charlottesville, VA, USA
[3]University of Chicago, Chicago, IL, USA

Corresponding Author:
John Levi Martin, University of Chicago, Sociology Department, 1126 E. 59th Street, Chicago, IL 60637, USA
Email: jlmartin@uchicago.edu

a set of moving window regressions over the network, to estimate the variance in such associations. The logic is simple and straightforward, the capacity to shed light on data high, and the limitations and drawbacks clear. We go into each of these in turn.

Approach

Our approach involves computing a local regression for each individual in the network, producing a coefficient that might correspond to her best guess as to the association of the variables in question, a guess based on her observations of those within her personal horizon. We will call all of those within her horizon her "neighborhood." We then wish to examine the distribution of all these local regression coefficients, to be able to characterize an overall network as having high or low variation across position. To formalize, let G be a network with a set of nodes N, each of which is observed on two variables, x and y. (We discuss the reason for this limitation to bivariate analyses below.) We are interested in the relation of these variables, as quantified by a regression slope. We favor a regression as opposed to a correlation because it aids in comparability of magnitude across local regressions, given that the variances of the coefficients will change from one local neighborhood to another. For the ith individual, let $Q(i)$ denote all the neighbors of i. We discuss below some of the ways that the investigator may construct this neighborhood. For all the members of $Q(i)$, which we will index by j, we may also construct a weight w_{ij} indicating the strength of the relation between i (the focal node) and j (some neighbor).[1] We discuss below some of the ways that the investigator may construct these weights. Thus for the ith individual, we fit the model

$$y_j = \beta_i x_j w_{ij} + c_i, \tag{1}$$

$$j \in Q(i). \tag{2}$$

The number of cases for the ith individual's regression is thus $|Q(i)|$ (and not the ith observation alone). The global model can be understood as a special case in which $Q(i) = N$ and $w_{ij} = 1$ for all i and j.

This is, as the alert reader will notice, an approach that is formally identical to that used in spatial analysis under the name geographically weighted regression (GWR) (see Fotheringham, Brunsdon, and Charlton, 2002). Just as with GWR, we construct $|N|$ local estimates of our slope parameter and are interested in the degree to which, and the pattern by which, this parameter varies over our data. We discuss the relevance of well-known limitations of GWR below.

We adapt the logic, however, for the case of networks, especially when considering how to define neighborhoods and how to define weights.

Definition of the Neighborhood

One way to define the neighborhood is to include all nodes within some distance of the focal node. By "distance" between two nodes, we mean the length of the shortest path in a network between them. If we denote this path length as $L(i, j)$ and choose some distance d, we may define a function D that indicates all the nodes within this distance of any focal node: $D(i, d) = \{j \mid L(i, j) \leq d\}$. We can then use this function to define our neighborhoods; thus $Q(i) = D(i, d)$. One might be interested in the special case of the simple neighborhood in which $d = 1$, and hence $Q(i)$ is all those nodes j to which i is tied ($Q[i] = \{j \mid x_{ij} = 1\}$). However, in most social networks, this neighborhood is too small to allow us to produce stable regression estimates.

This method will, however, usually lead the neighborhoods will vary in size across the graph. This can lead to some regression slopes to be based on many cases and others on few cases, which can confound volatility of our estimates with the variation we are interested in. For this reason, we may seek to hold constant across neighborhoods not the maximum distance but the number of neighbors to be included for each focal node (call this number M). To do this, we first find the smallest d that $D(i, d) > M$. We then include in $Q(i)$ all $D(i, d-1)$ and then determine how far short we are of M ($M - |D[i, d-1]|$); call this m^*. We then randomly select m^* nodes that are at distance d from i, producing a constant set of M of i's nearest neighbors.

A complication may arise if the graph G is disconnected, that is, if there are some pairs of nodes between which there is no connecting path. A subset of a graph that is connected is known as a "component"; within any component, all path lengths are finite, but between components, path lengths may be seen as infinite. If we are constructing neighborhood by looking for the M closest neighbors, members of components of a size less than M cannot have properly defined neighborhoods. However, it is worth emphasizing, that because we may also use weights based on the distance of neighbors of i from i, in some cases we may prefer to use what we shall call "unrestricted" neighborhoods (thus every neighborhood includes all nodes). This approach may be especially attractive where there are many small components.

Distance Weighting

However we compose our neighborhood, we have the possibility of weighting all members equally, or weighting them by

[1] We are, of course, free to consider $Q(i) = N$ for all i, only with $w_{ij} = 0$ for certain cases. In other words, rather than exclude some j from i's neighborhood, we set the weight between the two to zero. However, for compatibility with previous work, we make a distinction between the definition of the neighborhood of the ith node and the weights of the members these neighborhoods for the focal node.

some function of their closeness to *i*. Such weighting allows us to use an unrestricted approach to neighborhoods, which has the advantage of maximizing the available degrees of freedom for each local regression. Thus although constructing a neighborhood according to a fixed *d* or a fixed *M* may, if either of these is relatively small, often lead to unidentifiable models for neighborhoods in which there is no variation on the dependent or independent variable, here unidentification is unlikely to occur for nodes that are part of a large component.[2] For this reason, in our illustrations below, we use distance weighting combined with an unrestricted neighborhood.

Two commonly used functions for constructing these weights are the Gaussian and the bisquare. The bisquare function is

$$w_{ij} = \left[1 - \left(\frac{L(i,j)}{b}\right)^2\right]^2, \quad (3)$$

where b is a tunable parameter, which here we set to L^*, where L^* is the maximum observed path length (also known as the diameter of the graph).[3] We treat the path length between members of unconnected components as infinite and hence their weight as zero; alternatively, one can follow another common practice and set the distance between nodes in different components to be $L^* + 1$.

We are then interested in the *variation* of the set of local regression slopes, the vector $\beta = [\beta_1, \beta_2, \ldots \beta_N]$. There are two ways that we can quantify this variation. One involves using the variance, perhaps turning it into a standard deviation or a coefficient of variation (the variance divided by the mean). The second is to look at something like the interquartile range, which will be less sensitive to the presence of outliers than would a measure based on the standard deviation. Given that the purpose of this technique is to explore variation across local environments, we expect that analysts will prefer to keep the neighborhoods small, or the distance penalty in our weighting relatively severe, even at the cost of some extreme estimates, and so we recommend the latter approach.

However, even if all the individuals in the network were actually a random draw from a single, unstructured bag, we would normally expect some variation across local coefficients to arise merely because of the random allocation of respondents onto a network. To determine whether the observed variation is greater than that expected under chance sampling, we use a permutation test. We construct a number of simulated networks, in which we keep the overall structure the same as that of the observed network but randomly assign the persons to nodes. We can then examine where the observed variation (whatever measure we use) sits on this constructed distribution and produce a number that can be interpreted as a *p* value—how often we might expect this degree of variation or even more variation simply given the distribution of persons on the variables and given the structure of the network.

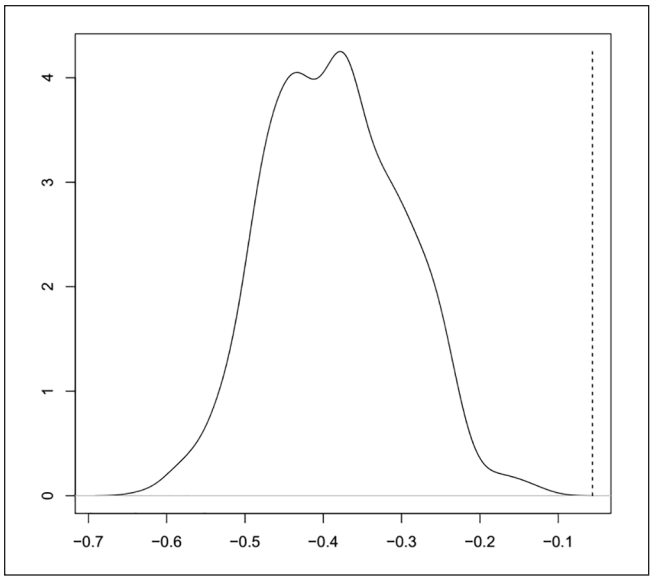

Figure 1. Density of local regression coefficients in one school.

Illustration

Observed Variation in Local Slopes

We here examine a number of schools from the National Longitudinal Study of Adolescent to Adult Health (Add Health) data set. We begin by presenting an example from a high school with 576 students with valid network data (out of 625 students altogether), in which we regress a scale of subjective feelings of being "connected" to the school on self-identification as Hispanic. In the data set as a whole, across all schools, the slope for this regression is −0.056, meaning that Hispanic students are somewhat less likely to feel connected, on average, than are non-Hispanic students. Figure 1 displays a smoothed density plot for β for this school. The dashed line indicates the aforementioned global regression slope across all schools in the data set. Given that the scale for "connectedness" has a standard deviation of 0.846, a local regression slope of 0.423, which is close to the mean local slope in this example school, implies that Hispanics are half a standard deviation less connected than non-Hispanics. Thus in some, but not all, parts of this school's network, the effect is rather substantial.

We also can, to a limited extent, visualize where in the network the slope is high and where it is low. Figure 2 assigns each node a shade on the basis of its local regression value.

[2]However, if there are isolated small components, such as dyads (sets of two nodes and their relations) and triads (sets of three nodes and their relations), we may have, under some weighting schemes, unidentified local regressions. We discuss solutions below.

[3]It is also possible empirically select *b* to maximize certain fit statistics.

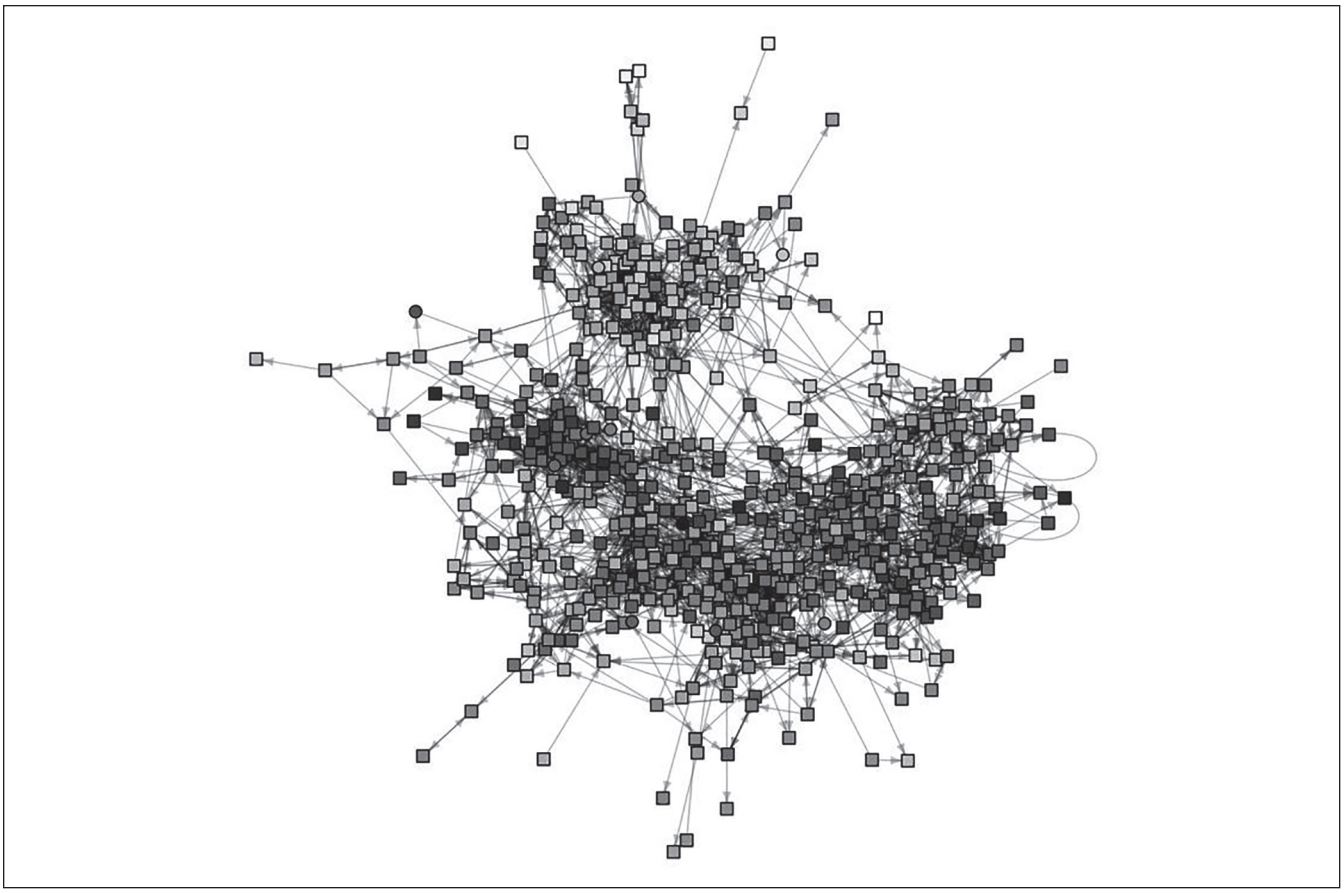

Figure 2. Network of school friendships and local effects.

Darker nodes are less negative than the lighter nodes. Hispanic students are indicated with a circle and non-Hispanic students with a square. Nodes are positioned here using the Fruchterman-Reingold algorithm.

We can see that there are two large clusters in the school, with relatively few ties between clusters. In the smaller one, there is a more negative relation between being Hispanic and feeling connected to the school. But there is also variation within the components: for example, in the larger component, the relationship between being Hispanic and disconnection is smallest in an area to the upper left.

We can compare such variances with those produced by the same analysis in a different school. Thus Table 1 compares the results given above ("school A," row 1, column 1) with those of a somewhat smaller school ("school B," column 2). Each value is the normed interquartile range of the slope coefficient from the bivariate regression specified in each row. We see that the school A has more variance in the relation between feelings of connection and being Hispanic than does school B. Table 1 presents two other rows corresponding to two other bivariate analyses. Both of these have the student's indegree, taken as a proxy for popularity, as the dependent variable. The first of these regresses indegree on students' estimated GPAs and the second on sex. The two schools have similar degrees of dispersion of local coefficients for the former, while school B is more dispersed on the latter than school A. Thus one network may have greater variation than a second network regarding one relationship and less variation on another.

Table 1. Comparison across Schools and Regressions.

	School A	School B
Regression		
Connected on Hispanic	7.36	6.69
	[.68]	[.65]
Indegree on GPA	3.39	3.67
	[.63]	[.04]
Indegree on Female	8.87	12.68
	[.80]	[.85]
n	576	415

Note: All variances are multiplied by 10^3; permutation test results are in brackets, scaled so that a larger number indicates less expected under chance.

Figure 3 summarizes these results in a way similar to that used in Figure 1. We gain additional insight by seeing that it is not simply that the relation between feelings of connection

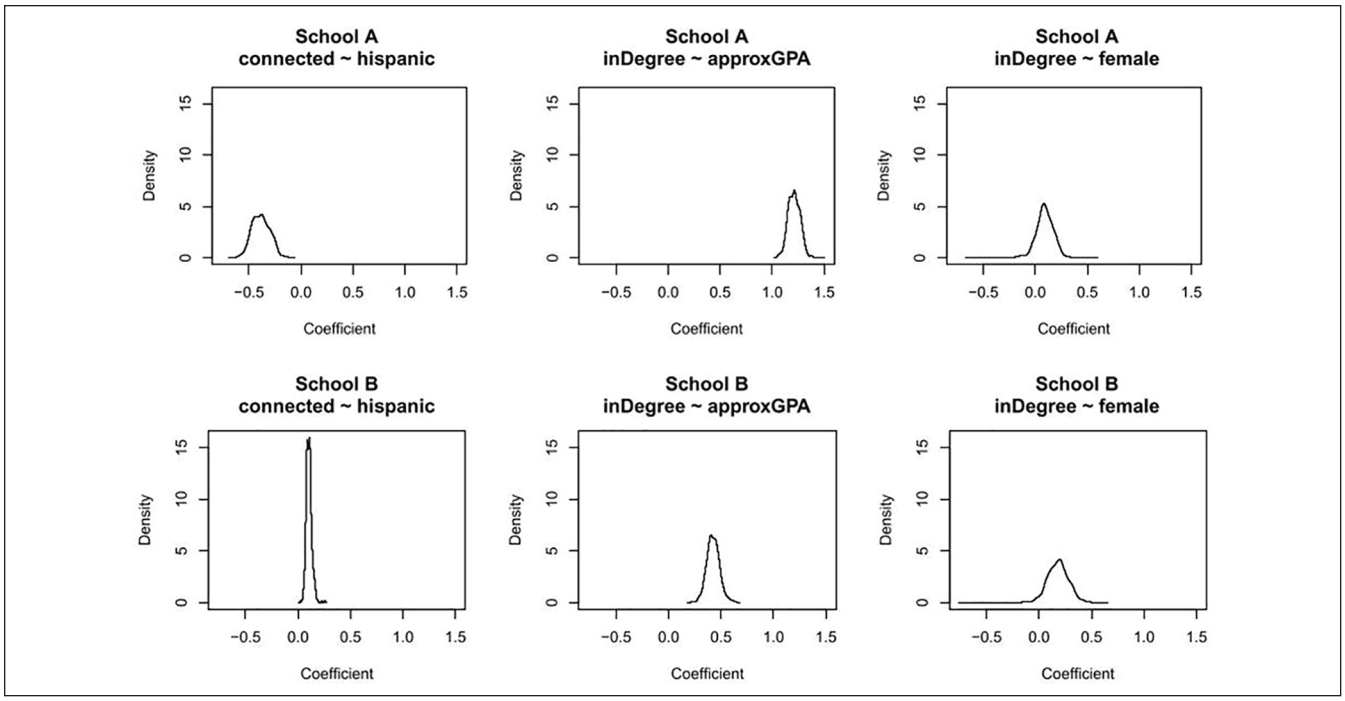

Figure 3. Comparisons of results in Table 1.

and identification as Hispanic is more concentrated in school B than in school A; the distributions of the two schools are on opposite sides of zero. Furthermore, we find that although the variation of the relation between indegree and GPA is similar in the two schools, it is substantially larger on average in school A than in school B.

Significance of Network Structure

In the example given in Figure 1, we saw substantial variation in the local coefficients linking students' identifying as Hispanic to their feeling of connectedness. However, there are two reasons that we might see such a variation in local slope parameters. On one hand, this pattern could be expected given the distribution of individuals on the dependent and independent variables. This does not mean that the variation is an artifact; it may well describe the phenomenological texture of the school's relational environment. However, we may be particularly interested in cases in which the variation has to do with the specific network structure; the variation, then, is a network property above and beyond the distribution of the individuals on the two variables in question.

To determine this, we can compare the degree of observed variation with that expected under a constructed distribution. In this case, we take the observed respondents but randomly assign them to different positions in the network structure. We then compute the distribution of local parameters for this constructed network and then a measure of the degree of variation, such as the variance or the interquartile range. For our focal example, the results from 100 such simulations find 32 with an interquartile range as great as that observed. This suggests that although there is some reason to think that the degree of local variability of the relation between Hispanic and connectedness has something to do with the social organization of this particular school, we are not confident that this degree of variation is really a network characteristic, as opposed to a characteristic of the set of individuals in the school.

In contrast, the relation between popularity and sex is, compared with such a constructed probability distribution, relatively more dispersed in both schools than is the relation between connection and Hispanic status.[4] (Table 1 includes these results for each regression in brackets.) Thus the permutation test facilitates within-case, but across-model, comparisons. Finally, it is interesting that although the total degree of variance in the relation between popularity and GPA is similar across the two schools (row 2), in the second school, we are not at all surprised to see such a relation given the individual distributions, whereas in the first, there is more evidence of a particular network effect.

[4]In this constructed distribution, we treat indegree as a proxy for popularity and leave it fixed as an individual attribute for reasons of expositional clarity. That is, we do not recompute it in each constructed distribution. We are thus constructing counterfactual worlds in which "popular people" may have few friends.

Discussion: Toward Multivariate Statistics

We have demonstrated the utility of local network regressions as a way of exploring heterogeneity in structural relationships in network data, an issue that is increasingly considered key in comparing dynamics within and across networks (e.g., Flashman 2012, 2014; McFarland et al. 2014). We have, it will be noted, examined only bivariate regressions. This is for an important reason: local regressions of this sort easily induce false correlations between independent variables in a multiple regression. This has been discovered for the case of GWRs (Páez, Farber, and Wheeler 2011; Wheeler and Tiefelsdorf 2005) and occurred in our own multivariate simulations. We therefore propose this form of local network regressions only for bivariate relations. However, we close by making a few tentative suggestions for ways of moving toward multivariate analyses.

One possibility would certainly be to move toward multilevel modeling in which the level 2 units are the nonnested neighborhoods of nodes. It is not, however, yet clear whether the distributional assumptions which are innocuous for conventional nested data structures would be problematic here.

A second possibility is to make use of the spatial filtering approach shown by Griffith (2008) to perform well in disentangling local effects in two-dimensional spatial problems. The problem with the direct application of spatial techniques to network data is that the weights matrix **W**, which automatically has certain advantageous properties in a metric space, may lead to imaginary solutions or no solutions outside of such a context. It would therefore make sense to use the network data to first position nodes in a space (which may be of high dimensionality) and then to use distance in this space to construct a weights matrix (which may also allow the application of techniques such as that suggested by Wheeler and Waller 2009 or Congdon 2006).

Such future explorations may or may not allow the robust identification of local effects from multivariate regressions. However, in any case, bivariate local network regressions are extremely promising exploratory and diagnostic tools that are simple to perform and to interpret.

Acknowledgments

We are grateful to reviewers and to the editors for comments that greatly increased the cogency of this contribution.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research uses data from Add Health, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris and funded by a grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu). No direct support was received from grant P01-HD31921 for this analysis.

References

Congdon, Peter. 2006. "A Model for Non-parametric Spatially Varying Regression Effects." *Computational Statistics & Data Analysis* 50(2):422–45.

Flashman, Jennifer. 2012. "Academic Achievement and its Impact on Friend Dynamics." *Sociology of Education* 85(1):61–80.

Flashman, Jennifer. 2014. "Friend Effects and Racial Disparities in Academic Achievement." *Sociological Science* 1:260–76.

Fotheringham, A. Stewart, Chris Brunsdon, and Martin Charlton. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Hoboken, NJ: John Wiley.

Griffith, Daniel A. 2008. "Spatial-Filtering-Based Contributions to a Critique of Geographically Weighted Regression (GWR)." *Environment and Planning A* 40(11):2751–69.

McFarland, Daniel A., James Moody, David Diehl, Jeffrey A. Smith, and Ruben J. Thomas. 2014. "Network Ecology and Adolescent Social Structure." *American Sociological Review* 79(6):1088-1121.

Páez, Antonio, Steven Farber, and David Wheeler. 2011. "A Simulation-Based Study of Geographically Weighted Regression as a Method for Investigating Spatially Varying Relationships." *Environment and Planning A* 43(12):2992–3010.

Wheeler, David, and Michael Tiefelsdorf. 2005. "Multicollinearity and Correlation among Local Regression Coefficients in Geographically Weighted Regression." *Journal of Geographic Systems* 7(2):161–87.

Wheeler, David C., and Lance A. Waller. 2009. "Comparing Spatially Varying Coefficient Models: A Case Study Examining Violent Crime Rates and Their Relationships to Alcohol Outlets and Illegal Drug Arrests." *Journal of Geographical Systems* 11(1):1–22.

Author Biographies

Peter McMahan is an assistant professor of sociology at McGill University. His research centers on the ways that communication informs the structural and cultural features of groups in varied social settings. He specializes in statistical modeling and computational methodologies, with particular focus on network analysis, natural language processing, and emerging methods of inference on large, unstructured data sets.

Adam Slez is an assistant professor in the Department of Sociology at the University of Virginia. Using spatial data analysis along with more traditional forms of historical inquiry, his recent work examines the way in which policies related to the expansion of the

railroad network contributed to patterns of third-party mobilization in the American West over the course of the late nineteenth century. His previous work on state and party formation at the U.S. Constitutional Convention of 1787 has appeared in the *American Sociological Review*.

John Levi Martin teaches sociology at the University of Chicago. He is the author of *Social Structures, The Explanation of Social Action, Thinking through Theory, Thinking through Methods*, and *Thinking through Statistics*, as well as articles on methodology, cognition, social networks, and theory.