Original Article

Methodological Innovations
Volume 9: 1–10
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2059799115622746
mio.sagepub.com

Log-linear distance models of homophily in small groups

Peter J Carrington

Abstract

This article demonstrates the innovative use of the log-linear distance model in the assessment of homophily in a set of small groups, such as the co-participants in a set of events. It traces the development of the application of the log-linear distance model to the study of homophily and reviews its recent use to assess the extent and structure of gender and age homophily in groups of criminal accomplices ('co-offenders'). The transformations of the group membership data that are a prerequisite of the log-linear analysis, and the interpretation of the log-linear parameters, are explained in detail in order to make the approach accessible to potential users. Although the described applications are to gender and age homophily in groups of criminal accomplices, the method can be used to assess homophily on one or more variables in any small groups.

Keywords

Log-linear, distance model, homophily, co-offending

Homophily – the tendency to associate with others who are similar to oneself – is ubiquitous in social relations (McPherson et al., 2001). Recently, the log-linear distance model has been used innovatively to model homophily on ordinal and dichotomous variables in small groups of criminal accomplices (Carrington, 2015a, 2015b); however, the methods are equally applicable to any small groups. Although the methods are introduced in the cited sources, their explanation of the methods is very brief. This article explains the methods in detail. It discusses the conceptualisation and operationalisation of homophily, traces the development of the application of the log-linear distance model to homophily and its advantages over the well-known Coleman index, shows how it can be applied to small groups and reviews the recent applications (Carrington, 2015a, 2015b). No new results are reported: the purpose of this article is to put the recent applications in conceptual and historical context and to make the model accessible to potential users.

Conceptualising homophily

Homophily has been defined as 'the principle that a contact between similar people occurs at a higher rate than among dissimilar people' (McPherson et al., 2001: 416). Mayhew et al. (1995) characterise homophily as 'the social distance interaction hypothesis': 'The frequency with which people enter into face-to-face association (contact, interaction) is inversely related to the social distance between them' (p. 20). Thus, according to the principle, or hypothesis, of homophily, the highest rate of interaction is expected between people who are in the same category of a discrete variable (e.g. both male) or who have the same value on a continuous variable (e.g. same age), that is, whose social distance is zero.

As the definition of McPherson et al. (2001) implies, the phenomenon of observed homophily, or homogeneity in social relationships, does not in itself imply any particular explanation. In particular, it is not necessarily due to personal choice (Mayhew et al., 1995: 20). Explanations of observed homophily generally fall into three categories:

1. *Population composition*: This might be characterised as the 'null' hypothesis. The selection of interaction partners is influenced simply by the availability of potential partners. If partners are selected at random, then their characteristics will reflect the distribution of those characteristics in the population or pool of potential partners. If the population is distributed unevenly over categories or values of some attribute, then

Department of Sociology and Legal Studies, University of Waterloo, Waterloo, ON, Canada

Corresponding author:
Peter J Carrington, Department of Sociology and Legal Studies, University of Waterloo, Waterloo, ON N2L 3G1, Canada.
Email: pjc@uwaterloo.ca

– assuming random selection of partners –members of the numerical majority group(s) will be more likely to select similar partners, and homophily will be observed (Blau, 1977).[1] McPherson et al. (2001: 419) refer to this as *baseline homophily*. For example, because the great majority of criminals are male, random (i.e. gender-blind) selection of accomplices would, *ceteris paribus*, produce a high proportion of homophilous male co-offending groups (and a smaller number of mixed-gender and all-female groups), and that is precisely what has been observed in numerous studies of co-offending (van Mastrigt and Carrington, 2014).

2. *Preference*: For various reasons, people *prefer* to associate and collaborate with those who are similar to them. According to McPherson et al. (2001), 'the psychology literature has demonstrated experimentally that attraction is affected by perceived similarity …' (p. 435). Similarly, in the area of professional collaboration, Steffensmeier and Terry (1986) concluded on the basis of interviews with male thieves that they preferred same-gender accomplices because they were perceived as more reliable and trustworthy.
3. *Social structure*: Features of the social structure apart from overall population composition may constitute 'opportunity structures' that favour homophilous contact and collaboration. For example, McPherson et al. (2001) point out that 'organizational foci', such as neighbourhood play groups, schools, workplaces and voluntary organisations, bring together people who are similar in age, ethnicity, socioeconomic status and sometimes gender (pp. 431–434). Felson (2003) describes 'convergence settings', where local youth 'hang out' and have the opportunity to meet and screen potential criminal accomplices. Such 'opportunity [to meet and screen] structures' are often partly or entirely segregated by age, gender, ethnicity and so on, thus restricting the pools of potential partners to similar people. Waring (2002), Schwartz et al. (2015) and others (reviewed in van Mastrigt and Carrington, 2014) have noted that it is the 'local' availability and accessibility of potential criminal partners, rather than their distribution in the larger population, that affect partner selection and that local availability and accessibility are structured by social networks based on family, friendship, neighbourhood and so on. Pettersson (2003) shows that residential ethnic segregation leads to ethnic segregation (homophily) in co-offending, although individual *preferences* may tend towards ethnic heterogeneity.

As the 'baseline' homophily that can be accounted for by random selection and population composition is often of little theoretical or substantive interest,[2] some researchers restrict the definition and study of homophily to that which exceeds expectations under random selection – sometimes termed *inbreeding bias* or *inbreeding homophily* (Coleman 1958; Currarini et al., 2009; Currarini and Redondo, 2011; Fararo and Sunshine, 1964; Laumann and Pappi, 1976: 55; McPherson et al., 2001: 419; Marsden, 1988; van Mastrigt and Carrington, 2014). For example, Coleman (1958: Appendix B) conceptualises homophily in sociometric choices as the extent to which the number of actual choices of persons in the same category as the chooser exceeds the number expected under random selection. Coleman's (1958: 36) index of homophily is category-specific

$$h_i = \frac{\left(a_{ii} - e_{ii}\right)}{\left(m_i - e_{ii}\right)} \tag{1}$$

where h_i is the (inbreeding) homophily of persons in category i, a_{ii} is the number of same-category choices made by members of category i, m_i is the total number of choices made by persons in category i and e_{ii} is the expected number of same-category choices by persons in category i, under random selection, which is equal to m_i multiplied by the proportion of persons of category i in the population.

Limitations of the Coleman index

Although the Coleman index of inbreeding homophily has been used extensively (for recent applications, see, for example, Currarini et al., 2009; Currarini and Redondo, 2011; van Mastrigt and Carrington, 2014), it has several limitations:

- It is based on a binary conceptualisation of similarity: persons chosen are either of the same category or not. This can be seen in formula (1). Therefore, its application is limited to dichotomous variables or to polytomous variables that can be analysed as dichotomies. The notion of social distance – other than 'same/different' – is unknown.
- The estimate of homophily that it produces is category-specific. No overall estimate of homophily in the population is available.
- Its sampling distribution is unknown, so there is no known standard error and therefore no inferential statistics such as confidence intervals or significance tests.
- It is not readily incorporated into a multivariate model, in which, for example, homophily can be assessed while controlling other relevant variables.

The log-linear distance model

The limitations of the Coleman index of homophily do not apply to the log-linear distance model. It applies to ordinal (and dichotomous) variables, it has a well-understood sampling distribution and therefore a standard error, from which confidence intervals and significance tests can be derived, and it is part of the statistically well-developed and well-understood log-linear model and, more generally, of the generalised linear model (Agresti, 2013: Chapters 9–11).

The simplest log-linear distance model is applied to a square cross-tabulation **H**(*i, j*) of pair-wise interactions, or dyads, in which the rows and columns represent the categories of the same ordinal (or dichotomous) attribute, and the cells (*i, j*) contain counts of the number of dyads comprising an actor in category *i* and an actor in category *j*. The model can be written as

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^X + \beta d_{ij} \qquad (2)$$

where μ_{ij} is the count (i.e. frequency of dyads) in cell (*i, j*), *X* is the row and column variable, indexed by *i* (rows) and *j* (columns) and $d = |i - j|$.

Like all conventional log-linear models for two-way tables, the distance model includes an intercept parameter λ, parameterising the overall frequency,[3] and row and column parameters λ_i^X and $\lambda_{j.}^X$ parameterising the overall (marginal) frequency of each row and column, thus capturing the effects of population composition on interaction frequencies (Agresti, 2013: 340).[4] In addition, the distance model includes the parameter β, which captures the effect of social distance: in this case, the distance – that is, difference – between categories of the variable. Thus, the β parameter captures the strength and direction of the association between attribute-based social distance and the frequency of interactions while controlling for population composition. Inbreeding homophily is indicated by cell counts that decrease for cells at increasing distances from the main diagonal of the cross-tabulation. As the d_{ij} term measures the distance between categories *i* and *j*, a positive estimate of β indicates that cell frequencies increase with increasing differences between *i* and *j* – that is, heterophily – and a negative estimate indicates homophily.

Development

The log-linear distance model was introduced by Goodman (1972, 1979: model 7, p. 810; see also Breiger, 1981; Haberman, 1974)[5] for square ordinal social mobility tables, in which cell counts (*i, j*) indicate the number of father–son dyads for which father is in the *i*th occupational class and son is in the *j*th occupational class. Cells on the main diagonal – where $i=j$ and father and son are in the same occupational class – indicate social immobility ('status inheritance'; i.e. inheritance by a son of his father's occupational class); cells at increasing distances from the main diagonal indicate increasing amounts of upward or downward social mobility. The β parameter (in model 2 above) captures the strength of the association between the number of occupational classes ($|i-j|$) separating father and son and the number of father–son dyads in cell (*i, j*), that is, the extent to which social mobility exceeds expectations based on random placement of sons in occupational classes.

Marsden (1988) applied the log-linear distance model to analyse homophily in personal networks. The 1985 American General Social Survey captured data on the characteristics of up to five persons with whom respondent 'discussed important matters'. Marsden (1988) transformed the data into (up to) five respondent–alter dyads for each respondent and then constructed cross-tabulations – on age, education, race/ethnicity, religion and gender – with rows representing respondents, columns representing alters and counts of dyads in the cells. Various forms of the log-linear distance model were applied to these cross-tabulations in order to assess the amount and structure of homophily on each attribute of respondents' choices of discussion partners.

Extension to small groups

The log-linear distance model applies to a cross-tabulation of dyadic interactions: the categories of the variable to which the two members of the dyad belong are indexed in the rows and columns of the cross-tabulation. Therefore, it is inapplicable to interactions involving more than two actors and therefore to groups of more than two members.[6] However, any 'small' or 'face-to-face' group incorporates a set of dyadic, or pair-wise, interactions or relationships: a group of size *n* generates $n(n-1)/2$ dyads (Mayhew et al., 1995: 25–27). This phenomenon is limited to 'small' groups because in larger groups it is unrealistic to assume that all members interact with all other members (Mayhew et al., 1995: 25). As Schaefer (2012) writes, in justifying the omission of larger groups from a study of social versus spatial distance in criminal collaborations,

> The assumption is made that individuals who were part of the same offense [i.e. co-offending group] had a social relationship. Two offenses were dropped from the data because they had such a high number of participants (23 and 54) that they likely violated this assumption. (p. 143)

Stegbauer and Rausch (2012) showed how lists of the members of a set of small groups of varying sizes can be used to generate a cross-tabulation of co-participating dyads, without knowledge of the identities of the members. Specifically, they transformed a list of the nationalities of the co-participants in a set of events (namely, sessions at an academic conference) into a set of co-participating dyads, labelled by the nationalities of the members, which in turn were used to generate a cross-tabulation of the frequencies with which people of different (and the same) nationalities co-participated in conference sessions. As nationality is a polytomous nominal variable, the concept of social distance was inapplicable, as was the log-linear distance model.

Carrington (2015b) extended the method of Stegbauer and Rausch (2012) to analyse age homophily among co-offenders (accomplices in crimes) apprehended by police in connection with all recorded co-offences in Canadian police during 2006–2009. Like Stegbauer and Rausch (2012), Carrington (2015b)

(a)

Co-offence	Co-offenders	Age
A	Jack	3
	Joe	3
	Harry	5
	Walter	6
	Mac	7
B	Jill	4
	Jerry	4
C	Mel	5
	Jo	6

(b)

M		Column	Age				
		Index	1	2	3	4	5
Row	Index	Label	3	4	5	6	7
Co-offence	1	A	2	0	1	1	1
	2	B	0	2	0	0	0
	3	C	0	0	1	1	0

(c)

M^TM		Column	Age				
		Index	1	2	3	4	5
Row	Index	Label	3	4	5	6	7
Age	1	3	4	0	2	2	2
	2	4	0	4	0	0	0
	3	5	2	0	2	2	1
	4	6	2	0	2	2	1
	5	7	2	0	1	1	1

(d)

H		Column	Age				
		Index	1	2	3	4	5
Row	Index	Label	3	4	5	6	7
Age	1	3	1	0	1	1	1
	2	4	0	1	0	0	0
	3	5	1	0	0	1	0.5
	4	6	1	0	1	0	0.5
	5	7	1	0	0.5	0.5	0

Figure 1. Transformation of co-offending data into an age-by-age matrix: (a) co-offending data, (b) weighted incidence matrix **M**, (c) weighted adjacency matrix $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ and (d) weighted adjacency matrix with corrected cell counts **H**.
Total dyad count = 12.

began with information on one characteristic of each of the co-participants in a set of events: specifically, the age in years of each of the offenders. The distance model was particularly appropriate because age (unlike nationality in the research of Stegbauer and Rausch, 2012) is an ordered attribute and was coded in the data as an ordinal variable, with values from 3 to 88 years. The co-offences in the data ranged in size (i.e. number of recorded participants) from 2 to 44 co-offenders, but the great majority had only 2 or 3 co-offenders.

Following the method of Stegbauer and Rausch (2012), Carrington (2015b) created a square symmetric cross-tabulation with 86 rows and columns, representing ages from 3 to 88 inclusive. Cell counts (*i, j*) indicated the number of co-offending dyads in which a person aged $i+2$ co-offended with a person aged $j+2$. Formally, the age-by-age co-offending matrix is

$$\mathbf{H} = \left(h_{ij}\right), \text{for :}$$
$$i, j \; = \; 1 \text{ to } 86 \text{ (with row and column labels 3 to 88 years), and}$$
$$h_{ij} \; = \; h_{ji} = \text{the number of co-offending dyads whose members have ages } i+2 \text{ and } j+2$$

The matrix **H** was constructed from the individual co-offences, which were transformed into a weighted incidence matrix **M**

$$\mathbf{M} = \left(m_{ij}\right), \text{for:}$$
$$i \; = \; 1 \text{ to } 443{,}056 \text{ co-offences, and}$$
$$j \; = \; 1 \text{ to } 86, \text{ for participations of offenders aged 3 to 88, and}$$
$$m_{ij} \; = \; \text{the number of participations in co-offence } i \text{ of offenders of age } j+2$$

For example, Figure 1(a) shows three hypothetical co-offences labelled A, B and C, involving nine co-offenders aged 3–7 years. The first co-offence has five co-offenders, of whom two are 3 years old, and one each is 5, 6 and 7 years old. This co-offence comprises 10 co-offending dyads: Jack–Joe, Jack–Harry, Jack–Walter, Jack–Mac, Joe–Harry and so on The second and third co-offences have only two co-offenders each and therefore comprise only one co-offending dyad each; in the second, they are both 4 years old, and in the third, they are 5 and 6 years old. Figure 1(b) shows the resulting rows of the incidence matrix **M**.

When **M** is pre-multiplied by its transpose, the result is a square matrix $\mathbf{M}^T\mathbf{M}$ (Figure 1(c)), showing the number of times that offenders of age $i+2$ co-offended with offenders of age $j+2$; for example, in the example in Figure 1, two 3-year-olds co-offended with one 5-year-old in co-offence A and none in co-offence B or C, so the value for $\mathbf{M}^T\mathbf{M}_{1,3}$ (and for $\mathbf{M}^T\mathbf{M}_{3,1}$) is 2. This adjacency matrix is symmetric, so the upper (or lower) triangle is redundant. However, the diagonal cells of $\mathbf{M}^T\mathbf{M}$ (shaded in Figure 1(c)) have incorrect values of 4, 4, 2, 2, 1 for numbers of co-offences between offenders of the same age. Stegbauer and Rausch (2012: 4) correct the diagonal entries with the formula

$$h_{ij} = \frac{\sum_k \left(m_{ki} \cdot \left(m_{kj} - 1 \right) \right)}{2} \quad \text{for } i = j \text{ and } k = 1, \ldots, s$$

where s is the number of rows in **M**, which in the example gives the correct diagonal entries of 1, 1, 0, 0, 0 (Figure 1(d)), indicating the single co-offence between 3-year-olds (in co-offence A; see Figure 1(a)) and between 4-year-olds (in co-offence B) and no same-age co-offences between 5-, 6- and 7-year-olds. A further correction was required in the Carrington (2015b) research, as the log-linear models were applied to the entire symmetric matrix, unlike the analysis in Stegbauer and Rausch (2012), where only the lower triangle was analysed. In $\mathbf{M^TM}$, each off-diagonal dyad is double-counted: once in the lower triangle and once in the upper triangle. In order to eliminate the double-counting, the modified formula used by Carrington (2015b) also divides the off-diagonal cell counts by 2

$$h_{ij} = \frac{\sum_k \left(m_{ki} \cdot m_{kj} \right)}{2} \quad \text{for } i \neq j \text{ and } k = 1, \ldots, s$$

where s is the number of rows in **M**, giving the cell counts shown in Figure 1(d), which sum to the correct total of 10 dyads.

The age-by-age matrix **H** in Carrington (2015b) contains 970,084 co-offending dyads: 176,503 dyads where both offenders are of the same age in years (i.e. on the main diagonal) and 793,581 dyads involving offenders of different ages. The distribution of ages of co-offending dyads is strongly affected by the row and column marginals – that is, the numbers of co-offenders of each year of age – which range from 31 dyads that include an 88-year-old to 106,696 dyads that include a 19-year-old. The effect of population composition - that is, the row and column marginal effects - can be removed by taking the normalized residuals from the log-linear independence model (Agresti, 2013: 339)[7]:

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^X \tag{3}$$

where μ_{ij} is the count (i.e. frequency of interactions) in cell (i, j) and X is the row and column variable, indexed by i and j.

The marginals can be plotted in an age-by-age matrix (Figure 2). In Figure 2, the density of the shading of cells is proportional to the normalised residual. As predicted by the homophily hypothesis, co-offending is greater than expected (from population composition) on and near the main diagonal (red cells) and less than expected off the diagonal (green cells). This is confirmed by the distance parameter estimate β from the log-linear distance model (formula (2)), which was −0.1398 (standard error (SE)=0.0002, $p<0.0001$): that is, negative and statistically significant (Carrington, 2015b: 345–346).

Further hypotheses based on age-by-age cross-tabulation were tested in Carrington (2015b), primarily concerning variations in homophily with co-offenders' ages – which can be seen in Figure 2, where homophily appears to be stronger among teenagers and children and practically non-existent among older co-offenders. This phenomenon was modelled by the 'variable distance' model, which allows the effect of social distance to vary

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^X + \beta_1 d_{ij} + \beta_2 d_{ij} s_{ij} \tag{4}$$

where $d=|i-j|$ and $s=i+j$ (Carrington, 2015b: formula 2, p. 345).

The β_2 parameter represents an interaction between heterophily and age; a positive estimate for β_2 indicates that age heterophily increases – that is, age homophily decreases – with an increase in the sum of the ages of the members of the dyad. Carrington (2015b) also tested hypotheses and models that added social distance based on age groups to distance based on year of age. The concept of 'age group' and models for it are discussed in the following section.

Extension to two variables

Carrington (2015a) extended the approach of Carrington (2015b) to model homophily on two variables (gender and age) simultaneously. Homophily on each variable was estimated while controlling for homophily on the other variable (as well as for population composition), and the interaction between homophily on the two variables was estimated: that is, whether or not co-offending between occupants of certain pairs of categories of the two variables occurred more or less frequently than expected on the basis of homophily on each of the variables individually.

Carrington (2015a) used the same data as Carrington (2015b), but the age range was trimmed to 5–75 years in order to have at least 10 female co-offenders for each year of age. The result was a reduction to 968,712 co-offending dyads. As with Stegbauer and Rausch (2012) and Carrington (2015b), the dyads were transformed into a cross-tabulation, but in this case it was a four-dimensional 71 × 71 × 2 × 2 cross-tabulation by age and gender simultaneously: $\mathbf{H}=(h_{ijkl})$, in which each cell h_{ijkl} is a count of the number of dyads in which the age of the first member is indexed by i and his or her gender by k, and the age of the other member is indexed by j and his or her gender by l. Formally, the cross-tabulation is denoted by

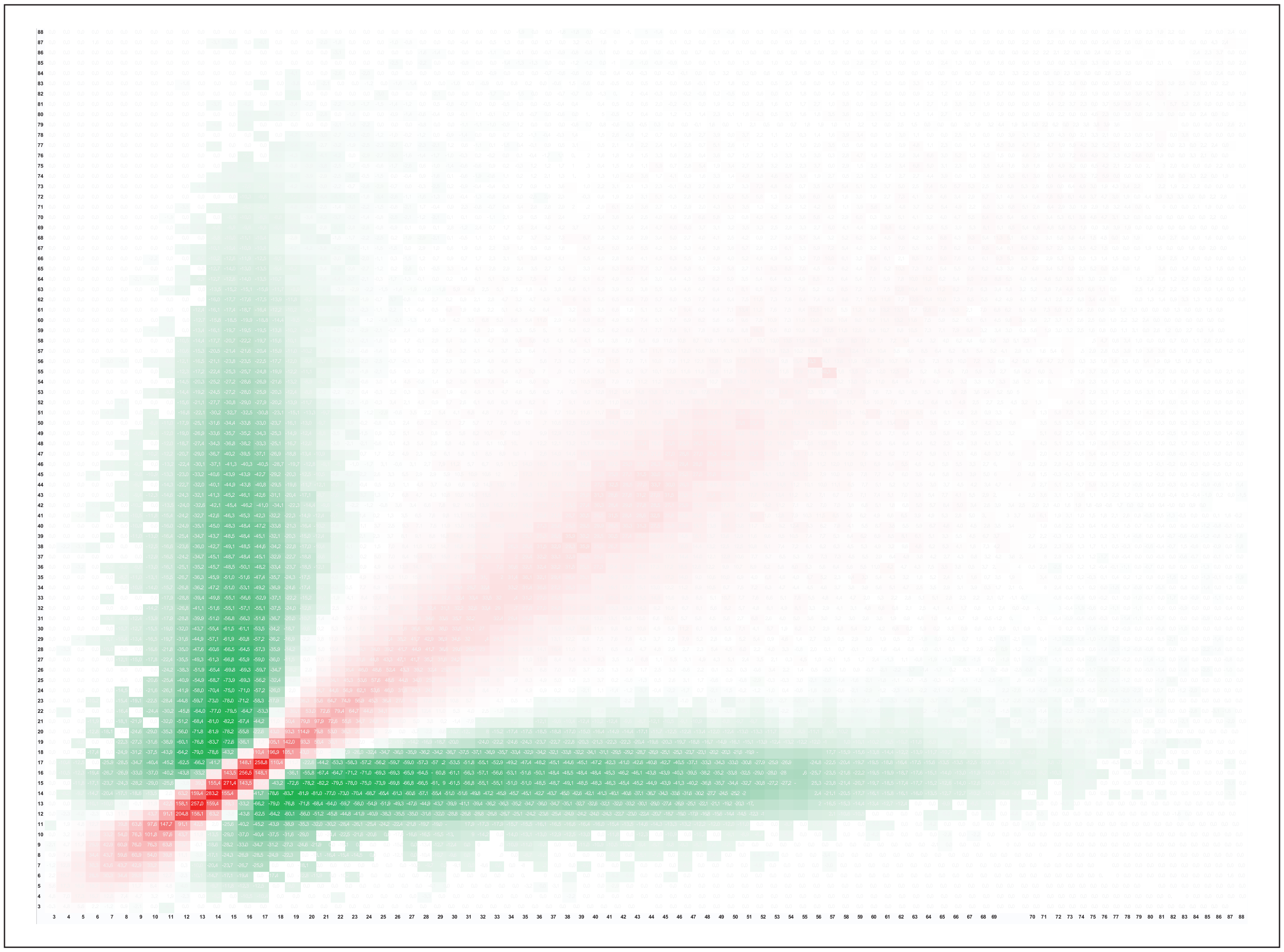

Figure 2. Age-by-age cross-tabulation with row and column marginal effects removed (residuals from the independence model). Negative residuals are shaded in green and positive residuals in red. Darkness of shading is proportional to the size of the residual. This figure will be available in colour in the online version.

$$\mathbf{H} = \left(h_{ijkl}\right)$$

where $i, j = \{1, 2, \dots 71\}$, labelled by ages $\{5, 6, \dots 75\}$; $k, l = \{1, 2\}$, labelled by {'Male', 'Female'}; and $h_{ijkl} = h_{jilk} \equiv$ the number of co-offending dyads whose members have the ages indexed by i and j and genders indexed by k and l.

To transform the individual co-offending groups into dyads, they were first transformed into two weighted incidence matrices **M** and **F**, for male and female co-offenders

$$\mathbf{M} = \left(m_{ij}\right)$$

where $i = \{1, 2, \dots, 442{,}534\}$, indexing co-offences; $j = \{1, 2, \dots, 71\}$, indexing participations of male offenders with ages 5, 6, …, 75; and m_{ij} = the number of participations in co-offence i of male offenders of age j.

$$\mathbf{F} = \left(f_{ij}\right),$$

where $i = \{1, 2, \dots, 442{,}534\}$, indexing co-offences, $j = \{1, 2, \dots, 71\}$, indexing participations of female offenders with ages 5, 6, … ,75, and f_{ij} = the number of participations in co-offence i of female offenders of age j.

For example, Figure 3(a) shows three co-offences labelled A, B and C, involving nine co-offenders aged 5–9 years. The first co-offence has five co-offenders, all male, of whom two are 5 years old and one each is 7, 8 and 9 years old. This co-offence generates 10 co-offending dyads. The second and third co-offences have only two co-offenders each and generate one dyad each; in the second co-offence, they are both 6 years old and of different genders, and in the third co-offence, they are 7 and 8 years old and of different genders. Figure 3(b) shows the resulting incidence matrices **M** and **F**.

When **M** is pre-multiplied by its transpose, the result is a square adjacency matrix $\mathbf{M}^{T}\mathbf{M}$ (Figure 3(c)), showing the number of times that male offenders of age i co-offended with male offenders of age j; for example, in the example in Figure 3, two 5-year-old males co-offended with one 7-year-old male in co-offence A and none in co-offence B or C, so the value for $\mathbf{M}^{T}\mathbf{M}_{1,3}$ (and for $\mathbf{M}^{T}\mathbf{M}_{3,1}$) is 2. However, the diagonal cells of $\mathbf{M}^{T}\mathbf{M}$ (shaded in Figure 3(c)) have incorrect values of 4, 1, 1, 2, 1 for numbers of co-offences between male offenders of the same age. The Stegbauer and Rausch

(a)

Co-offence	Co-offenders	Age	Gender	Age index	Gender index
A	Jack	5	Male	1	1
	Joe	5	Male	1	1
	Harry	7	Male	3	1
	Walter	8	Male	4	1
	Mac	9	Male	5	1
B	Jill	6	Female	2	2
	Jerry	6	Male	2	1
C	Sally	7	Female	3	2
	Joe	8	Male	4	1

(b)

M		Column	Age				
		Index	1	2	3	4	5
Row	Index	Label	5	6	7	8	9
Co-offence	1	A	2	0	1	1	1
	2	B	0	1	0	0	0
	3	C	0	0	0	1	0

F		Column	Age				
		Index	1	2	3	4	5
Row	Index	Label	5	6	7	8	9
Co-offence	1	A	0	0	0	0	0
	2	B	0	1	0	0	0
	3	C	0	0	1	0	0

(c)

$\mathbf{M}^{\mathbf{T}}\mathbf{M}$		Age index				
		1	2	3	4	5
Age	1	4	0	2	2	2
	2	0	1	0	0	0
	3	2	0	1	1	1
	4	2	0	1	2	1
	5	2	0	1	1	1

$\mathbf{M}^{\mathbf{T}}\mathbf{F}$		Age index				
		1	2	3	4	5
Age	1	0	0	0	0	0
	2	0	1	0	0	0
	3	0	0	0	0	0
	4	0	0	1	0	0
	5	0	0	0	0	0

$\mathbf{F}^{\mathbf{T}}\mathbf{M}$		Age index				
		1	2	3	4	5
Age	1	0	0	0	0	0
	2	0	1	0	0	0
	3	0	0	0	1	0
	4	0	0	0	0	0
	5	0	0	0	0	0

$\mathbf{F}^{\mathbf{T}}\mathbf{F}$		Age index				
		1	2	3	4	5
Age	1	0	0	0	0	0
	2	0	1	0	0	0
	3	0	0	1	0	0
	4	0	0	0	0	0
	5	0	0	0	0	0

(d)

$\mathbf{H}_{11}$		Age index				
		1	2	3	4	5
Age	1	1	0	1	1	1
	2	0	0	0	0	0
	3	1	0	0	.5	.5
	4	1	0	.5	0	.5
	5	1	0	.5	.5	0

$\mathbf{H}_{12}$		Age index				
		1	2	3	4	5
Age	1	0	0	0	0	0
	2	0	.5	0	0	0
	3	0	0	0	0	0
	4	0	0	.5	0	0
	5	0	0	0	0	0

$\mathbf{H}_{21}$		Age index				
		1	2	3	4	5
Age	1	0	0	0	0	0
	2	0	.5	0	0	0
	3	0	0	0	.5	0
	4	0	0	0	0	0
	5	0	0	0	0	0

$\mathbf{H}_{22}$		Age index				
		1	2	3	4	5
Age	1	0	0	0	0	0
	2	0	0	0	0	0
	3	0	0	0	0	0
	4	0	0	0	0	0
	5	0	0	0	0	0

Figure 3. Transformation of co-offending groups to cross-tabulated dyads: (a) co-offending data, (b) weighted co-offending incidence matrices, (c) adjacency matrices with uncorrected cell counts and (d) the four slices of the four-dimensional cross-tabulation, with corrected cell counts.
Total dyad count = 12.

(2012: 4) correction (see the previous section and the formula below) gives the correct diagonal entries of 1, 0, 0, 0, 0 (Figure 3(d)), indicating the single co-offence between male 5-year-olds (in co-offence A; see Figure 3(a)) and no same-age co-offences between male 6- to 9-year-olds:

$$m'_{ij} = \frac{\sum_k \left(m_{ki} \cdot \left(m_{kj} - 1\right)\right)}{2} \quad \text{for } i = j \text{ and } k = 1, 2, \ldots s$$

where s is the number of rows in **M**.

As in Carrington (2015b; see preceding section), a further correction was required, because the entire symmetric matrix was analysed, unlike the analysis in Stegbauer and Rausch (2012), where only the lower triangle was analysed. In $\mathbf{M}^{\mathbf{T}}\mathbf{M}$, each off-diagonal dyad is double-counted: once in the lower triangle and once in the upper triangle. In order to eliminate the double-counting and preserve the correct total number of co-offending dyads, the modified formula also divides the off-diagonal cell counts by 2

$$m'_{ij} = \frac{\sum_k \left(m_{ki} \cdot m_{kj}\right)}{2} \quad \text{for } i \neq j \text{ and } k = 1,2,\ldots,s$$

where s is the number of rows in **M,** resulting in the male*male slice of **H**, or $\mathbf{H}_{11} = (h_{ij11})$, where the last two subscripts, 1 and 1, index male gender (Figure 3(d)).

The same procedure was followed to produce the female*female co-offending adjacency matrix $\mathbf{H}_{22} = (h_{ij22})$ and the male*female co-offending matrix $\mathbf{H}_{12} = (h_{ij12})$ and its transpose, the female*male co-offending matrix $\mathbf{H}_{21} = (h_{ij21})$ (Figure 3(c) and (d)).

The 71 × 71 × 2 × 2 cross-tabulation of co-offending dyads was then reduced to a 3 × 3 × 2 × 2 cross-tabulation by aggregating years of age into *age statuses*, which Carrington (2015a) argues are more sociologically and criminologically relevant than merely chronological years of age or the *ad hoc* groupings of ages into 'age groups' in Carrington (2015b). Following Burt (1991), dyadic co-offending frequencies were row- and column-normalised using iterative proportional fitting (IPF),[8] in order to remove the effects of varying volumes of co-offending by members of different year-of-age groups. Distances (dissimilarities) between the co-offending patterns of each pair of years-of-age were calculated as the Euclidean distance between the corresponding pair of rows (or, equivalently, columns) in the two-dimensional 71 × 71 year-of-age by year-of-age cross-tabulation of

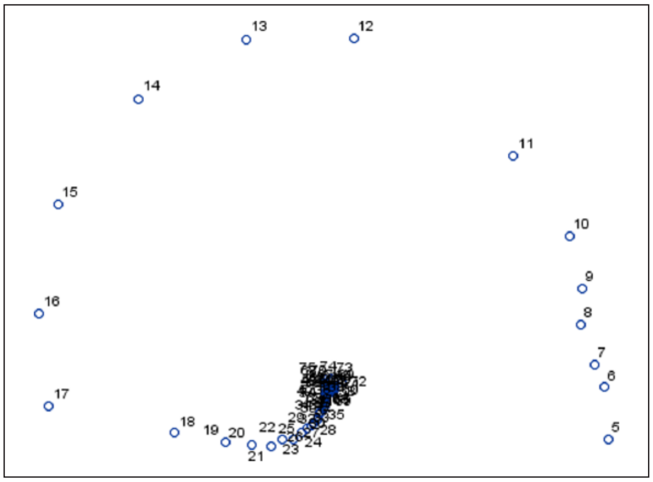

Figure 4. Two-dimensional MDS plot of distances between years-of-age, based on co-offending patterns.

normalised co-offending frequencies (Burt, 1991). Two methods were used to cluster the years of age in the resulting 71 × 71 distance matrix: visual examination of a multidimensional scaling (MDS) plot and Ward's hierarchical agglomerative clustering method.[9] For Ward's clustering, the choice of the number of clusters was assisted by plots of the pseudo-*F* statistic and pseudo-t^2 statistic against the number of clusters (Cooper and Milligan, 1988). Ward's clustering solution suggested either three clusters – 5–11, 12–17 and 18–75 years – or four clusters, with 18–75 years split into 18–45 and 46–75 years. The MDS plot (Figure 4) suggests (the same) three age clusters: 5–11, 12–17 and 18–75 years. Although derived from a clustering based on (dis-)similarities in co-offending patterns, these age clusters correspond exactly to the three age categories defined in Canadian criminal law: children (5–11 years), who cannot be prosecuted for criminal offences; youth (12–17 years), who are under the jurisdiction of the youth justice system; and adults, who are under the jurisdiction of the ordinary criminal justice system (Carrington, 2015a).[10]

The estimates for the gender and age status heterophily parameters were both negative and statistically significant (−0.757 and −1.897, respectively), indicating gender homophily and even stronger age status homophily among the co-offenders. The estimate for the interaction between gender and age status was 0.323 ($SE = 0.008$, $p < 0.0001$), indicating that dyads comprising a young female and an adult male are *more* frequent than expected from the main effects of gender and age, that is, gender and age status homophily are mutually attenuating, not reinforcing (Carrington, 2015a).

Limitations

The log-linear model of homophily in groups is a 'dyad-independent' model that assumes there are no connections between groups (Handcock et al., 2008). Where data are available on connections between groups – that is, if information is available on the presence of the same persons in two or more groups – then a model such as the exponential random graph model (ERGM) is, in principle, more appropriate because it can estimate homophily (using the same underlying distance model) while controlling for network effects (Robins and Daraganova, 2013: 93). However, there are limitations to the size of population that can be modelled with an ERGM (Robins and Lusher, 2013), whereas the approach described here is insensitive to population size, as the nodes are aggregated into categories of the classificatory variable.

The log-linear distance model also has the limitation that the transformation of groups into dyads, preparatory to construction of the cross-tabulation required by the log-linear model, loses information about group size. As homophily has been found to be related to group size (McPherson et al., 2001; Mayhew et al., 1995; van Mastrigt and Carrington, 2014), this may be an issue in some applications.

Conclusion

The log-linear distance model is a powerful and flexible way of modelling inbreeding homophily in small disconnected groups on attributes measured with ordinal or dichotomous variables. It brings the modelling of homophily into the overall family of log-linear models, whose statistical properties are well-known and whose parameterizations are extremely flexible and therefore readily adaptable to the testing of specific hypotheses. In order to apply the log-linear model to data on group memberships, the list of the members of each group must first be converted to a cross-tabulation of dyadic relationships, following the method of Stegbauer and Rausch (2012). Examples of the use of the log-linear distance model to assess age and gender homophily in co-offending groups are provided in recent work by Carrington (2015a, 2015b).

Acknowledgements

The author gratefully acknowledges the continuing support of the Canadian Centre for Justice Statistics, Statistics Canada for this programme of research and thanks Yvan Clermont, Julie McAuley, Anthony Matarazzo and Marian Radulescu for their assistance in accessing the data. The article has benefited from comments on earlier versions by Ronald Breiger, John Scott and an anonymous referee The opinions expressed herein are those of the author alone and do not represent the opinions of Statistics Canada.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The research was supported by a grant from the Social Sciences and Humanities Research Council of Canada.

Notes

1. Members of minority groups will also be more likely to select members of majority groups, creating observed heterophily,

but this will be less than the homophilous selections of the majority.

2. Although it can be consequential. For example, baseline homophily and the numerical predominance of 'Anglos' in the American population account for the more 'racially homogeneous' social networks of American Anglos (McPherson et al., 2001: 420). Also, a 'highly skewed' proportion of one type of person – such as males in the legitimate or criminal workplace – can lead to a (psychological) preference among that type for homophily and can create local social structures that promote homophily among members of that type (Kanter, 1977; McPherson et al., 2001; Steffensmeier, 1983).
3. Or the intercept may be omitted, in which case the overall frequency is incorporated into the row and column parameters.
4. In the general log-linear model of a two-way table, the row and column variables may be distinct (i.e. the table may be rectangular), so the parameter for the column variable is designated λ_j^Y (Agresti, 2013: 340), but the distance model is applied to a square table in which the rows and columns refer to the same variable.
5. Haberman (1974) terms this the 'fixed distance' model, to distinguish it from the 'variable distance' model, in which the effect of the difference between categories is allowed to vary (formula (4) below). For an example of that model, see Carrington (2015b).
6. In principle, a higher dimensional cross-tabulation could be used to capture interactions among more than two actors and a corresponding log-linear model to model homophily; however, the author is unaware of any such work.
7. This is the same as the distance model (formula (2)), except the distance (heterophily) parameter is omitted so that any homophily will appear in the residuals.
8. Using the IPF procedure in UCINET (Borgatti et al., 2002).
9. Using PROC MDS and PROC CLUSTER in SAS (SAS Institute Inc., 2013).
10. As Carrington (2015a) points out, these are practically the same groupings of ages that were arrived at by Carrington (2015b), using a different conceptualisation of 'age group' and correspondingly different methods.

References

Agresti A (2013) *Categorical Data Analysis*. Hoboken, NJ: Wiley.

Blau PM (1977) *Inequality and Heterogeneity*. New York: Free Press.

Borgatti SP, Everett MG and Freeman LC (2002) *Ucinet for Windows: Software for Social Network Analysis*. Harvard, MA: Analytic Technologies.

Breiger RL (1981) The social class structure of occupational mobility. *American Journal of Sociology* 87(3): 578–611.

Burt RS (1991) Measuring age as a structural concept. *Social Networks* 13(1): 1–34.

Carrington PJ (2015a) Gender and age segregation and stratification in criminal collaborations. *Journal of Quantitative Criminology*. Epub ahead of print 19 October. DOI: 10.1007/s10940-015-9269-2.

Carrington PJ (2015b) The structure of age homophily in co-offending groups. *Journal of Contemporary Criminal Justice* 31(3): 337–353.

Coleman J (1958) Relational analysis: The study of social organizations with survey methods. *Human Organization* 17: 28–36.

Cooper MC and Milligan GW (1988) The effect of measurement error on determining the number of clusters in cluster analysis. In: Gaul W and Schader M (eds) *Data, Expert Knowledge and Decisions*. Berlin: Springer, pp. 319–328.

Currarini S and Redondo FV (2011) *A simple model of homophily in social networks*. Research Paper Series No. 24. Department of Economics, University Ca' Foscari of Venice. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1697503

Currarini S, Jackson MO and Pin P (2009) An economic model of friendship: Homophily, minorities and segregation. *Econometrica* 77: 1003–1045.

Fararo TJ and Sunshine MH (1964) *A Study of a Biased Friendship Net*. Syracuse, NY: Youth Development Center, Syracuse University.

Felson M (2003) The process of co-offending. In: Smith MJ and Cornish DB (eds) *Theory for Practice in Situational Crime Prevention*. Monsey, NY: Criminal Justice Press, pp. 149–167.

Goodman LA (1972) Some multiplicative models for the analysis of cross-classified data. In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press, pp. 649–696.

Goodman LA (1979) Multiplicative models for the analysis of occupational mobility tables and other kinds of cross-classification tables. *American Journal of Sociology* 84(2): 804–818.

Haberman SJ (1974) *The Analysis of Frequency Data*. Chicago, IL: University of Chicago Press.

Handcock MS, Hunter DR, Butts CT, et al. (2008) statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software* 24(1): 1–11.

Kanter R (1977) *Men and Women of the Corporation*. New York: Basic Books.

Laumann EO and Pappi FU (1976) *Networks of Collective Action: A Perspective on Community Influence Systems*. New York: Academic Press.

McPherson M, Smith-Lovin L and Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27: 415–444.

Marsden PV (1988) Homogeneity in confiding relations. *Social Networks* 10: 57–76.

Mayhew BH, McPherson JM, Rotolo T, et al. (1995) Sex and race homogeneity in naturally occurring groups. *Social Forces* 74(1): 15–52.

Pettersson T (2003) Ethnicity and violent crime: The ethnic structure of networks of youths suspected of violent offences in Stockholm. *Journal of Scandinavian Studies in Criminology and Crime Prevention* 4: 143–161.

Robins G and Daraganova G (2013) Social selection, dyadic covariates, and geospatial effects. In: Lusher D, Koskinen J and Robins G (eds) *Exponential Random Graph Models for Social Networks*. Cambridge: Cambridge University Press, pp. 91–100.

Robins G and Lusher D (2013) What are exponential random graph models? In: Lusher D, Koskinen J and Robins G (eds) *Exponential Random Graph Models for Social Networks*. Cambridge: Cambridge University Press, pp. 9–15.

SAS Institute Inc. (2013) *SAS/STAT® 13.1 User's Guide*. Cary, NC: SAS Institute Inc.

Schaefer DR (2012) Youth co-offending networks: An investigation of social and spatial effects. *Social Networks* 34: 141–149.

Schwartz J, Conover-Williams M and Clemons K (2015) Thirty years of sex stratification in violent crime partnerships and groups. *Feminist Criminology* 10: 60–91.

Steffensmeier DJ (1983) Organization properties and sex segregation in the underworld: Building a sociological theory of sex differences in crime. *Social Forces* 6: 1010–1032.

Steffensmeier DJ and Terry RM (1986) Institutional sexism in the underworld: A view from the inside. *Sociological Inquiry* 56(3): 304–323.

Stegbauer C and Rausch A (2012) How international are international congresses? *Connections* 32(1): 1–11.

van Mastrigt SB and Carrington PJ (2014) Sex and age homophily in co-offending networks: Opportunity or preference? In: Morselli C (ed.) *Crime and Networks*. Abingdon: Routledge, pp. 28–51.

Waring EJ (2002) Co-offending as a network form of social organization. In: Waring E and Weisburd D (eds) *Crime and Social Organization*. New Brunswick, NJ: Transaction Publishers, pp. 31–47.

Author biography

Peter J Carrington is Professor of Sociology and Legal Studies at the University of Waterloo. His current research project, the Canadian Criminal Careers and Criminal Networks Study, combines his long-standing interests in social network analysis and in the development of crime and delinquency. His articles have appeared in various journals, including *Criminology*, *Journal of Quantitative Criminology*, *Journal of Mathematical Sociology*, *Social Networks* and *Canadian Journal of Criminology and Criminal Justice*. He is editor of *Applications of Social Network Analysis* (Sage Publications, 2014) and co-editor of *The SAGE Handbook of Social Network Analysis* (Sage Publications, 2011) and *Models and Methods in Social Network Analysis* (Cambridge University Press, 2005).