FIXED-EFFECT REGRESSIONS ON NETWORK DATA

Koen Jochmans* Sciences Po, Paris

Martin Weidner‡ University College London

May 24, 2017

Abstract

This paper studies inference on fixed effects in a linear regression model estimated from network data. An important special case of our setup is the two-way regression model, which is a workhorse method in the analysis of matched data sets. Networks are typically quite sparse and it is difficult to see how the data carry information about certain parameters. We derive bounds on the variance of the fixed-effect estimator that uncover the importance of the structure of the network. These bounds depend on the smallest non-zero eigenvalue of the (normalized) Laplacian of the network and on the degree structure of the network. The Laplacian is a matrix that describes the network and its smallest non-zero eigenvalue is a measure of connectivity, with smaller values indicating less-connected networks. These bounds yield conditions for consistent estimation and convergence rates, and allow to evaluate the accuracy of first-order approximations to the variance of the fixed-effect estimator. The bounds are also used to assess the bias and variance of estimators of moments of the fixed effects.

Keywords: fixed effects, graph, Laplacian, network data, two-way regression model, variance bound, variance decomposition.

JEL classification: C23, C55

*Sciences Po, Département d'économie, 28 rue des Saints Pères, 75007 Paris, France. E-mail: koen.jochmans@sciencespo.fr.

‡University College London, Department of Economics, Gower Street, London WC1E 6BT, United Kingdom, and CeMMAP. E-mail: m.weidner@ucl.ac.uk.

We are grateful to Ulrich Müller and three referees for constructive comments. We would also like to thank Bryan Graham, Áureo de Paula, Valentin Verdier, and, in particular, Jean-Marc Robin for stimulating discussions and for insightful comments on earlier drafts of this paper. Jochmans gratefully acknowledges financial support from the European Research Council through Starting Grant nº 715787. Weidner gratefully acknowledges support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001 and from the European Research Council grant ERC-2014-CoG-646917-ROMIA. The first version of this paper dates from August 4, 2016 and is available at https://arxiv.org/abs/1608.01532v1.

1 Introduction

Data on the interaction between economic agents are in increasing supply. Matched data, where the interaction between two types of agents is observed, are one type of such network data. Examples here include data that link workers to firms and students to teachers. Such data can be used to study a variety of issues. Abowd, Kramarz and Margolis (1999) investigate the sign and magnitude of assortative matching between workers and firms. Card, Heining and Kline (2013) analyze the contribution of worker and firm heterogeneity to the variance of earnings. Finkelstein, Gentzkow and Williams (2016) perform a similar decomposition of health care utilization. Rockoff (2004) and Aaronson, Barrow and Sander (2007) assess teacher effectiveness in the classroom.

The workhorse method used in all these studies is a linear regression model with fixed effects. While such regressions are well studied in the conventional panel data context, applications to networks are different for at least two reasons. Firstly, it are often the fixed effects and their moments (rather than some slope coefficient that is common across observations) that are of primary interest. Secondly, the structure of network data is quite different from that of panel data. In typical applications, it will be quite sparse. In the context of teacher effectiveness, for example, students necesarilly interact with only few teachers and the size of any classroom is bounded from above. This implies that fixed effects will be estimated from few observations, even though the total size of the data set may be quite substantial. Relying on large-sample approximations may then be highly misleading.

While it is intuitive that the network structure will be an important determinant of the accuracy of statistical inference, the structure of a network becomes complex rather fast. This implies that it is difficult to see how the data carry information about certain parameters. In this paper we analyze this issue in a linear version of the Bradley and Terry (1952) model on a general network; this setup encompasses the two-way fixed-effect model for linked data. We do so by acknowledging that the data structure of a network can be translated to a graph where agents are vertices and an edge between vertices is

present if agents interact. The usefulness of graph theory in establishing existence of the fixed-effect estimator was noted in Abowd, Creecy and Kramarz (2002) in the context of matched employer-employee data. Here, we go beyond this observation to study the statistical precision of the estimator of the fixed effects and of plug-in estimators of their moments.

We show that the variance of the fixed-effect estimator is equal to the inverse of the Laplacian matrix of the network. Like the adjacency matrix, this matrix fully summarizes the network. However, the Laplacian considers the connectivity of the vertices in the network. A bound on the variance of the fixed-effect estimator is obtained that depends inversely on the smallest non-zero eigenvalue of the (normalized) Laplacian. This eigenvalue is a measure of connectivity, with smaller values indicating less-connected networks.[1] The larger it is, the more dense is the network. One interesting consequence of this bound is that consistent estimation is possible even if the network becomes less connected as the sample grows.

We also refine the variance bound to uncover how the local structure of the network around a given vertex influences the variance of the vertex-specific parameter estimator. Clearly, the variance of such an estimator is decreasing in the degree of the vertex — which is the number of neighbors of the vertex, and equals the total number of observations for that vertex specific parameter. In addition, the improved bounds account for the sensitivity of the variance with respect to the degree of the neighbors of the vertex, thus sharpening the estimator's convergence rates.

Sampling noise in the fixed-effect estimator translates into bias in the estimator of their moments. Rockoff (2004), for example, notes that the sample variance of his estimated fixed effects will tend to overestimate the true variance of the teacher effects. Our variance bounds on the fixed effects readily yield bounds on the bias and variance of estimators of their moments. These bounds again depend on the smallest non-zero eigenvalue of the

[1] Eigenvalues and eigenvectors of network matrices have also been found of use in determining equilibrium conditions in games on networks (Bramboullé, Kranton and D'Amours, 2014) and in (statistical) community detection (Schiebinger, Wainwright and Yu, 2015).

(normalized) Laplacian matrix, as well as on various weighted harmonic means of the degree sequences of the network. Consistent estimation turns out to be possible only in sufficiently dense networks. Even then, bias correction will be needed for inference to be size correct.

In Section 2 we introduce the basic version of the model and estimator under study. In Section 3 we present our bounds. In Section 4 we discuss various extensions of our baseline model. In Section 5 we discuss weighted graphs. Proofs and additional results are available as supplementary material.

2 Model and estimator

Consider a graph $\mathcal{G} := \mathcal{G}(V, E)$ where $m := |E|$ edges are placed between $n := |V|$ vertices. For the largest part of the paper we will work with a simple undirected graph, without loops (i.e., no edge connects a vertex with itself). Without loss of generality we label the vertices by natural numbers, so $V = \{1, \dots, n\}$. The set E contains the $m \leq n(n-1)/2$ unordered pairs (i, j) from the product set $V \times V$ that are connected by an edge, where we assume throughout that $m > 0$. Vertices i and j are said to be connected if $\mathcal{G}$ contains a path from i to j, and the graph $\mathcal{G}$ is said to be connected if every pair of vertices in the graph is connected. We will work under the convention that $i < j$ for $(i, j) \in E$. Our analysis is invariant to this choice of orientation.

2.1 A fixed-effect model

Our interest lies in estimating a linear regression model where outcomes are labelled by elements of E. The simplest form of our setup is as follows. For each $(i, j) \in E$ we observe the outcome

$$y_{ij} = \alpha_i - \alpha_j + u_{ij}, \qquad u_{ij} \sim \text{i.i.d. } N(0, \sigma^2), \tag{2.1}$$

where $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ are vertex-specific parameters to be estimated and the $u_{ij} \in \mathbb{R}$ are unobserved disturbances with unknown variance σ^2. For notational convenience in the sequel it is useful to define $y_{ji} = -y_{ij}$ and $u_{ji} = -u_{ij}$. Equation (2.1) is overparametrized,

so we impose that

$$\sum_{i=1}^{n} \alpha_i = 0. \tag{2.2}$$

Other normalizations on the α_i could be chosen but (2.2) will prove convenient for our purposes. A normalization can be dispensed with if the object of interest is changed to parameter differences, i.e, to $\alpha_i - \alpha_j$. Corresponding results for such differences are given in the supplementary material.

In the complete graph, where $m = n(n-1)/2$, deriving the sampling behavior of the maximum-likelihood estimator of the α_i in (2.1) is rather standard. Here we are interested in incomplete graphs. To motivate this we first recall the concept of a bipartite graph.

Example 1 (Bipartite graph)**.** Partition V as $V_1 \cup V_2$ and consider a bipartite graph. That is, suppose that E is a subset of the product set $V_1 \times V_2$. Then edges are formed between the vertex sets V_1 and V_2 but not within V_1 and V_2. So, for an edge (i, j) we necessarily have that $i \in V_1$ and $j \in V_2$. In this case, our model (2.1) can be obtained from the specification

$$y_{ij} = \mu_i + \eta_j + u_{ij} \tag{2.3}$$

by setting

$$\alpha_i = \begin{cases} \mu_i & \text{if } i \in V_1, \\ -\eta_i & \text{if } i \in V_2. \end{cases}$$

Choosing the sign in front of η_i is without loss of generality because links are only formed between, but never within, V_1 and V_2. □

Example 1 is a stripped-down version of the classical regression model with two-way fixed effects. This is a workhorse model to capture unobserved heterogeneity across units in matched data. One leading example are value-added models for measuring the effectiveness of teachers in the classroom (Rockoff, 2004). Here, the (i, j) represent student-teacher pairs, and the graph $\mathcal{G}$ is far from complete. Indeed, each student is only taught by a handful of teachers and the size of a classroom cannot increase without bound. Therefore, even though the data set may be very large, student and teacher effects are estimated from very small subsamples.

Relying on large-sample theory in an incomplete graph may be unwarranted. We focus on variance bounds that are valid for small samples and shed light on how these bounds depend on the structure of $\mathcal{G}$. The model in (2.1) is convenient for these purposes, as it allows to most easily convey our main points. Once this has been done we will show that introducing edge-specific covariates or allowing for non-normal and heteroskedastic errors does not alter our main findings.

We will also show how our results extend to weighted graphs. One important situation where such graphs arise is with panel data, that is, when we observe multiple outcomes for each $(i, j) \in E$. This is frequently the case with linked data sets and, in fact, variation across time may be needed to be able to disentangle parameters. In student-teacher data it may allow to separate teacher effects from classroom effects. The wage regressions of Abowd, Kramarz and Margolis (1999) are another case in point. There, the (i, j) represent pairs of workers and firms, and workers need to be employed at different firms over time to separately estimate worker and firm effects. Even then, the worker effects will be difficult to estimate precisely due to limited mobility of workers over time (Andrews, Gill, Schank and Upward 2008) and one may decide to focus on the estimation of firm effects. A weighted graph naturally arises as a consequence of this profiling out, as we show in the next example.

Example 1 (cont'd) (Bipartite graph)**.** As before, consider the two-way fixed-effect model

$$y_{ij} = \mu_i + \eta_j + u_{ij}$$

for a bipartite graph on $V = V_1 \cup V_2$. Suppose that our main focus is on the parameters η_j. For each $i, j \in V_2$, let $[i, j] := \{k \in V_1 : (k, i) \in E \text{ and } (k, j) \in E\}$. Then $m_{ij} := |[i, j]|$ is the number of vertices in V_1 that connect to both i and j. Differencing the above equation gives

$$y_{ijk} = \eta_i - \eta_j + u_{ijk}, \qquad k = 1, \ldots, m_{ij},$$

where $y_{ijk} := y_{ki} - y_{kj}$ and $u_{ijk} := u_{ki} - u_{kj}$. Thus, the original two-way model can be translated to a setup as in (2.1) that involves only pairs of vertices from V_2 where, possibly, multiple outcomes are observed for each pair. □

2.2 Estimation and inference

Under our convention that $i < j$ for $(i,j) \in E$ the (oriented) incidence matrix of $\mathcal{G}$ is the $m \times n$ matrix $\boldsymbol{B}$ with entries

$$(\boldsymbol{B})_{ei} := \begin{cases} 1 & \text{if the } e^{\text{th}} \text{ edge is given by } (i,j) \in E \text{ for some } j \in V, \\ -1 & \text{if the } e^{\text{th}} \text{ edge is given by } (j,i) \in E \text{ for some } j \in V, \\ 0 & \text{otherwise.} \end{cases}$$

The incidence matrix fully describes $\mathcal{G}$. Note that the oriented incidence matrix is unique up to negation of any of the columns, since negating the entries of a row corresponds to reversing the orientation of an edge. Moreover, the analysis to follow is invariant to our choice of orientation. Indeed, changing the orientation of the edge (i,j) jointly with the sign of y_{ij} leaves model (2.1) invariant. Throughout, the network structure is treated as fixed, that is, $\boldsymbol{B}$ is conditioned on.

Let $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_n)'$. Collect all outcomes in the m-vector $\boldsymbol{y}$ and all regression errors in the m-vector $\boldsymbol{u}$. Write $\boldsymbol{\iota}_n$ for the n-vector of ones and $\boldsymbol{I}_m$ for the $m \times m$ identity matrix. Equations (2.1)–(2.2) can then be written as

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{u}, \qquad \boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_m), \qquad \boldsymbol{\alpha}'\boldsymbol{\iota}_n = 0.$$

Because of normality of $\boldsymbol{u}$, the maximum-likelihood estimator of $\boldsymbol{\alpha}$ is equal to the (ordinary) least-squares estimator, that is,

$$\widehat{\boldsymbol{\alpha}} := (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)' = \underset{\boldsymbol{a} \in \{\boldsymbol{a} \in \mathbb{R}^n : \boldsymbol{a}'\boldsymbol{\iota}_n = 0\}}{\arg\min} \ \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}\|^2, \tag{2.4}$$

where $\|\cdot\|$ denotes the Euclidean norm.

We first address existence and uniqueness of $\widehat{\boldsymbol{\alpha}}$. Here and later, we let $\boldsymbol{M}^\dagger$ denote the Moore-Penrose pseudoinverse of matrix $\boldsymbol{M}$.

Lemma 1 (Existence)**.** *Let $\mathcal{G}$ be connected. Then*

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{B}'\boldsymbol{B})^\dagger \boldsymbol{B}'\boldsymbol{y}$$

and is unique.

The need for a pseudoinverse arises because $\boldsymbol{B}'\boldsymbol{B}$ is singular, as $\boldsymbol{B}\boldsymbol{\iota}_n = 0$. The use of the Moore-Penrose pseudoinverse follows from our normalization choice on $\boldsymbol{\alpha}$, that is, $\boldsymbol{\alpha}'\boldsymbol{\iota}_n = 0$. The result of the lemma is intuitive and in line with results in the literature on matched employer-employee data (Abowd, Creecy and Kramarz 2002).

The following theorem is immediate.

Theorem 1 (Sampling distribution)**.** *Let $\mathcal{G}$ be connected. Then*

$$\widehat{\boldsymbol{\alpha}} \sim N(\boldsymbol{\alpha},\ \sigma^2\,(\boldsymbol{B}'\boldsymbol{B})^{\dagger})$$

for any n.

The main implication of Theorem 1 is that, for an $n \times r$ matrix $\boldsymbol{R}$ of maximal column rank that is linearly independent of $\boldsymbol{\iota}_n$,

$$\frac{m-(n-1)}{r}\,\frac{(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})'\boldsymbol{R}\,(\boldsymbol{R}'(\boldsymbol{B}'\boldsymbol{B})^{\dagger}\boldsymbol{R})^{-1}\boldsymbol{R}'\,(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})}{(\boldsymbol{y}-\boldsymbol{B}\widehat{\boldsymbol{\alpha}})'(\boldsymbol{y}-\boldsymbol{B}\widehat{\boldsymbol{\alpha}})} \sim F_{r,m-n+1}, \tag{2.5}$$

which can be used to test the null hypothesis that $\boldsymbol{R}\boldsymbol{\alpha} = \boldsymbol{0}$ against the alternative that $\boldsymbol{R}\boldsymbol{\alpha} \neq \boldsymbol{0}$.

While Theorem 1 ensures size-correct inference it does not aid in understanding when the F-statistic in (2.5) will have low power or when the corresponding confidence sets will be wide. In the sequel we aim to understand how the structure of the network affects the standard error of the least-squares estimator. Such an analysis also yields conditions for consistent estimation and asymptotically-valid inference under non-normality for sequences of growing networks.

3 Network structure and variance bounds

Theorem 1 shows that, up to the scalar factor σ^2, the variance of $\widehat{\boldsymbol{\alpha}}$ is completely determined by the $n \times n$ Laplacian matrix of $\mathcal{G}$,

$$\boldsymbol{L} := \boldsymbol{B}'\boldsymbol{B} = \boldsymbol{D} - \boldsymbol{A},$$

where $\boldsymbol{D} := \mathrm{diag}(d_1, \ldots, d_n) = \mathrm{diag}\,(\boldsymbol{B}'\boldsymbol{B})$ is the degree matrix and $\boldsymbol{A}$ is the $n \times n$ adjacency matrix of $\mathcal{G}$, with entries

$$(\boldsymbol{A})_{ij} := \begin{cases} 1 & \text{if } (i,j) \in E \text{ or } (j,i) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Note that d_i, the degree of i, equals the number of vertices that vertex i is connected to.

It will be convenient to work with the normalized Laplacian

$$\boldsymbol{S} := \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}} = \boldsymbol{I}_n - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}.$$

We have $(\boldsymbol{L}^\dagger)_{ii} = d_i^{-1}(\boldsymbol{S}^\dagger)_{ii}$, and so

$$\mathrm{var}(\widehat{\alpha}_i) = \mathbb{E}((\widehat{\alpha}_i - \alpha_i)^2) = \frac{\sigma^2}{d_i}\,(\boldsymbol{S}^\dagger)_{ii}. \tag{3.1}$$

Equation (3.1) highlights the importance of the degree d_i, which is the effective number of observations that are used to infer α_i. However, (3.1) does not imply that $\mathrm{var}(\widehat{\alpha}_i)$ shrinks as $d_i \to \infty$, nor would it give a convergence rate if it did, as the normalized Laplacian matrix of $\mathcal{G}$ also changes when n grows.

3.1 Zero-order bound

To make progress on bounding the variance, let λ_i denote the ith eigenvalue of $\boldsymbol{S}$, arranged in increasing order; so, $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. The following results on those eigenvalues are from Chung (1997, Lemma 1.7). We have $\min_i \lambda_i = 0$ and $\max_i \lambda_i \leq 2$. Zero is always an eigenvalue of $\boldsymbol{S}$ because $\boldsymbol{B}\boldsymbol{\iota}_n = 0$, but, if $\mathcal{G}$ is connected, it has multiplicity one. That is, $\lambda_2 > 0$ is the smallest non-zero eigenvalue of the normalized Laplacian when $\mathcal{G}$ is connected. For the complete graph $\mathcal{G}$ (i.e., when $m = n(n-1)/2$) we have $\lambda_2 = n/(n-1)$, and for any $\mathcal{G}$ that is not complete we have $\lambda_2 \leq 1$.

The following result bounds the variance of $\widehat{\boldsymbol{\alpha}}$ as a function of λ_2.[2]

[2]We equally have $\mathrm{var}(\widehat{\alpha}_i) \leq \sigma^2/\tilde{\lambda}_2$, where $\tilde{\lambda}_2$ is the smallest non-zero eigenvalue of the (unnormalized) Laplacian $\boldsymbol{L}$. The spectrum of $\boldsymbol{L}$ has been the subject of more study than that of $\boldsymbol{S}$ in the graph literature. However, $\tilde{\lambda}_2 \leq n/(n-1)\,\min_{i \in V} d_i$. Thus, $\tilde{\lambda}_2$ may be very small (and the corresponding bound on $\mathrm{var}(\widehat{\alpha}_i)$ very large) as soon as a single vertex in V has a small degree, making it an unattractive quantity for our purposes.

Theorem 2 (Global bound)**.** *Let $\mathcal{G}$ be connected. Then*

$$\operatorname{var}(\widehat{\alpha}_i) \leq \frac{1}{d_i}\frac{\sigma^2}{\lambda_2}.$$

To interpret the bound it is useful to connect it to the Cheeger constant,

$$C := \min_{U \in \{U \subset V : 0 < \sum_{i \in U} d_i \leq m\}} \frac{\sum_{i \in U}\sum_{j \notin U}(\boldsymbol{A})_{ij}}{\sum_{i \in U} d_i}.$$

The constant $C \in [0,1]$ reflects how difficult it is to disconnect $\mathcal{G}$ by removing edges. Moreover, a larger value of C implies a more strongly-connected graph. From Chung (1997, Theorems 2.1 and 2.3),

$$2C \geq \lambda_2 \geq 1 - \sqrt{1 - C^2} \geq \frac{1}{2}C^2. \tag{3.2}$$

Thus, $\lambda_2 \in (0, 1 + 1/n]$ is a measure of global connectivity of the graph, and Theorem 2 implies that inference will be more precise when the graph is more strongly connected.

Theorem 2 also allows to derive some asymptotic properties under sequences of growing networks $\mathcal{G}$. Firstly, we find the pointwise consistency result

$$(\widehat{\alpha}_i - \alpha_i) \xrightarrow{p} 0 \quad \text{if} \quad \lambda_2 \, d_i \to \infty.$$

This result allows $\lambda_2 \to 0$ as $n \to \infty$. Secondly, letting h be the harmonic mean of the sequence $d_1, \ldots, d_n$, we have

$$\frac{\mathbb{E}(\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|^2)}{n} \leq \frac{1}{h}\frac{\sigma^2}{\lambda_2},$$

and so

$$\frac{\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|}{\sqrt{n}} \xrightarrow{p} 0 \quad \text{if} \quad \lambda_2 \, h \to \infty \quad \text{as} \quad n \to \infty, \tag{3.3}$$

by an application of Markov's inequality.

Example 2 (Erdős-Rényi graph)**.** Consider the Erdős and Rényi (1959) random-graph model, where edges between n vertices are formed independently with probability p_n. The threshold on p_n for $\mathcal{G}$ to be connected is $\ln(n)/n$ (Hoffman, Kahle and Paquette 2013). That is, if $p_n = c\ln(n)/n$ for a constant c, then, as $n \to \infty$, with probability approaching one, $\mathcal{G}$ is disconnected if $c < 1$ and connected if $c > 1$. In the former case, $\lambda_2 \to 0$ while, in the latter case, $\lambda_2 \to 1$, almost surely. □

3.2 First-order bound

A refinement of Theorem 2 takes into account the connectivity of the direct neighbors of i. Here, we call a direct neighbor, or a path-one neighbor, a vertex to which i is connected via a path of length one. Similarly, we will call those vertices that have geodesic distance equal to two from i path-two neighbors of i. The collection of direct neighbors of vertex i is

$$[i] := \{j \in V : (i,j) \in E \text{ or } (j,i) \in E\};$$

note that $|[i]| = d_i$. Let

$$h_i := \left(\frac{1}{d_i} \sum_{j \in [i]} \frac{1}{d_j} \right)^{-1}, \tag{3.4}$$

the harmonic mean of the degrees of all $j \in [i]$. Note that, for a given vertex i, h_i is increasing in the degree of its direct neighbors.

Theorem 3 (First-order bound)**.** *Let $\mathcal{G}$ be connected. Then*

$$\frac{\sigma^2}{d_i}\left(1 - \frac{2}{n}\right) \leq \operatorname{var}(\widehat{\alpha}_i) \leq \frac{\sigma^2}{d_i}\left(1 - \frac{2}{n} + \frac{1}{\lambda_2 h_i}\right).$$

Theorem 3 states that, for a given degree d_i and global connectivity measure λ_2, the upper bound on the variance of $\widehat{\alpha}_i$ is smaller if the direct neighbors of vertex i are themselves more strongly connected to other vertices in the network. Another implication of the theorem is the rate refinement

$$\operatorname{var}(\widehat{\alpha}_i) = \frac{\sigma^2}{d_i} + o(d_i^{-1}), \tag{3.5}$$

provided that $\lambda_2 h_i \to \infty$ as $n \to \infty$. Furthermore, the parametric rate is achievable even if λ_2 is not treated as fixed.

In the supplementary material we present a refinement of Theorem 3 that accounts for the dependence on h_i in the lower bound as well, and also adjusts the upper bound for overlap between $[i]$ and the sets $[j_1], \ldots, [j_{d_i}]$ for $j_1, \ldots, j_{d_i} \in [i]$, that is, for common neighbors. These bounds can be particularly useful when h_i is small, but are vacuous when

all path-two neighbors of vertex i are also path-one neighbors. This is the case, for example, in the complete graph.

We illustrate the usefulness of improving on Theorem 2 in a random graph.

Example 2 (cont'd). Consider the Erdős and Rényi (1959) random-graph model with $p_n = c\ln(n)/n$ for $c > 1$. Let i be a randomly chosen vertex. Then, as $n \to \infty$, we have, almost surely,

$$\lambda_2 \to 1, \qquad \frac{d_i}{\ln n} \to c, \qquad \frac{h_i}{\ln n} \to c.$$

Consequently,

$$\operatorname{var}(\widehat{\alpha}_i) = \frac{\sigma^2}{d_i} + O(d_i^{-2})$$

follows from Theorem 3. □

Additional examples that illustrate our results in analytically-tractable cases where $\lambda_2 \to 0$ as $n \to \infty$ are provided in the supplementary material.

3.3 Moments

Moments of the α_i may be of interest. Plug-in estimators of such moments are biased, in general, and inconsistent in sparse graphs. Here, we provide results for the plug-in estimator

$$\widehat{\vartheta} := n^{-1}\sum_{i=1}^{n}\widehat{\alpha}_i^2$$

of the variance $\vartheta := n^{-1}\sum_{i=1}^{n}\alpha_i^2$, which is arguably the most popular such parameter.[3] Results for higher-order moments are available in the supplementary material.

The variance estimator is upward biased, which is intuitive. A small calculation gives

$$\mathbb{E}(\widehat{\vartheta} - \vartheta) = \sigma^2\frac{\operatorname{tr}(\boldsymbol{L}^\dagger)}{n}.$$

[3] Card, Heining and Kline (2013) decompose the variance of log wages to back out the contributions of worker and firm effects. Similarly, the contribution of the variance in teacher effects to variation in test scores, the teacher's value added (Rockoff, 2004), is a key parameter in the study of educational outcomes.

Thus, the magnitude of the bias is driven by the Laplacian of the graph. In the complete graph, for example, we have $\mathrm{tr}(\boldsymbol{L}^\dagger) = (n-1)/n$ so that the bias is of order n^{-1}. This is the usual bias order in nonlinear estimators. Simple bounds that are valid more generally are

$$\frac{\sigma^2}{h}\left(1-\frac{2}{n}\right) \leq \mathbb{E}(\widehat{\vartheta}-\vartheta) \leq \frac{\sigma^2}{h}\left(1-\frac{2}{n}+\frac{1}{\lambda_2\, H}\right),$$

where we introduce the weighted harmonic mean of the h_i,

$$H := \left(\sum_{i=1}^{n} \frac{(h/n)/d_i}{h_i}\right)^{-1}, \quad \text{with weights satisfying} \quad \sum_{i=1}^{n}(h/n)/d_i = 1.$$

These bounds reveal that, $\mathbb{E}(\widehat{\vartheta}-\vartheta) = \sigma^2/h + o(h^{-1})$ if $\lambda_2 H \to \infty$. Therefore, for the bias to vanish asymptotically we need that the degrees of the individual vertices grow with n for an increasing fraction of the vertices.

Moving on, under normality we have

$$\mathrm{var}(\widehat{\vartheta}) = 4\sigma^2\frac{\boldsymbol{\alpha}'\boldsymbol{L}^\dagger\boldsymbol{\alpha}}{n^2} + 2\sigma^4\frac{\|\boldsymbol{L}^\dagger\|^2}{n^2},$$

where $\|.\|$ is the Frobenius norm. A simple (but conservative) upper bound on this variance can be obtained that is $O(h^{-1})$ if ϑ is bounded and $\lambda_2 H$ is bounded away from zero. We thus have that

$$(\widehat{\vartheta}-\vartheta) \xrightarrow{p} 0 \quad \text{if } h \to \infty \text{ as } n \to \infty,$$

under the same condition on $\lambda_2 H$.

To assess whether the bias is important for the accuracy of inference we require a more precise rate for the variance. The order of the variance turns out to depend on the data generating process of the graph and of the fixed effects. In the supplementary material we provide regularity conditions under which $\mathrm{var}(\widehat{\vartheta})$ is at most of order $\lambda_2^{-1}/(nh)$ and at least of order $1/(nh)$. The bias will therefore dominate sampling noise unless $\lambda_2\, n/h$ or n/h, respectively, converges to a finite constant.[4] Note that n/h will tend to diverge in sparse

[4]If the α_i are draws from a distribution P and we wish to estimate $\theta := \int \alpha^2\, dP(\alpha)$, then $\vartheta - \theta$ will typically be of order $n^{-1/2}$, and so the noise induced by sampling from P dominates the noise in $\widehat{\vartheta}-\vartheta$ under our familiar condition $\lambda_2\, h \to \infty$. The asymptotic bias of order h^{-1} then dominates the standard deviation of order $n^{-1/2}$ if n/h^2 is not close to zero, which again will often be the case in sparse networks.

networks, so that $\widehat{\vartheta}$ may fail to have a properly centered limit distribution in that case. Even in more dense networks, where n and h grow at the same rate, bias correction will typically be needed for test statistics to have correct size.

4 Extensions

4.1 Introducing covariates

Let us augment (2.1) with a set of p edge-specific covariates $\boldsymbol{x}_{ij}$ (again with $\boldsymbol{x}_{ji} = -\boldsymbol{x}_{ij}$), which we will treat as fixed for simplicity of notation. Collecting the covariates in the $m \times p$ matrix $\boldsymbol{X}$ and denoting by $\boldsymbol{\beta}$ the associated vector of regression slopes our model becomes

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{u}, \qquad \boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_m).$$

By standard results on partitioned regression, the maximum-likelihood estimator of $\boldsymbol{\alpha}$ now is

$$\widetilde{\boldsymbol{\alpha}} := (\boldsymbol{B}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{B})^{\dagger}\boldsymbol{B}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}, \qquad \boldsymbol{M}_{\boldsymbol{X}} := \boldsymbol{I}_m - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}',$$

provided that the regressors are not perfectly collinear.[5] In contrast, the (now infeasible) estimator to which the results from the previous section apply assumes $\boldsymbol{\beta}$ to be known and is

$$\widehat{\boldsymbol{\alpha}} := (\boldsymbol{B}'\boldsymbol{B})^{\dagger}\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Theorem 4 bounds the difference in the variance between these two estimators in terms of

$$\rho := \left\| (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{B}}\boldsymbol{X} \right\|_2, \qquad \boldsymbol{M}_{\boldsymbol{B}} := \boldsymbol{I}_m - \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{\dagger}\boldsymbol{B}',$$

where $\|\cdot\|_2$ denotes the spectral norm. $\rho \in [0, 1]$ is a measure of non-collinearity between the columns of $\boldsymbol{X}$ and $\boldsymbol{B}$, with ρ close to zero indicating near-collinearity. Indeed, while $\boldsymbol{X}'\boldsymbol{X}$ measures the total variation in $\boldsymbol{X}$, $\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{B}}\boldsymbol{X}$ captures the residual variation in $\boldsymbol{X}$, after its linear dependence on $\boldsymbol{B}$ has been partialled out. To state the result we also introduce $\overline{\boldsymbol{x}}_i := d_i^{-1} \sum_{j \in [i]} \boldsymbol{x}_{ij}$, and $\boldsymbol{\Omega} := \boldsymbol{X}'\boldsymbol{X}/m$.

[5] Verdier (2016) provides limit theory for the estimator of $\boldsymbol{\beta}$ in two-way fixed-effect regressions.

Theorem 4 (Covariates)**.** *Let* $\mathcal{G}$ *be connected,* $\operatorname{rank}(\boldsymbol{X}) = p$*, and* $\operatorname{rank}((\boldsymbol{X}, \boldsymbol{B})) = p+n-1$*. Then*

$$|\operatorname{var}(\widetilde{\alpha}_i) - \operatorname{var}(\widehat{\alpha}_i)| \;\leq\; \frac{2\,\sigma^2}{\rho}\left(\frac{1-\rho}{\lambda_2\, d_i\, h_i} + \frac{\overline{\boldsymbol{x}}_i'\,\boldsymbol{\Omega}^{-1}\overline{\boldsymbol{x}}_i}{m}\right)$$

for all $i \in V$*.*

Thus, if ρ is bounded away from zero, the first term appearing in the bound is of the same order as the last term in Theorem 3, and the second term is $O(m^{-1})$. Consequently, all results on $\widehat{\alpha}_i$ from the previous section straightforwardly extend to $\widetilde{\alpha}_i$.

4.2 Non-normal and heteroskedastic errors

Inference based on Theorem 1 presumes normality of the errors. In contrast, the bounds in Theorem 3 continue to hold when the errors in (2.1) are non-normal, as the variance of $\widehat{\alpha}_i$ depends only on the first and second moments of the data. The asymptotic statements obtained in the previous section also carry over. We now discuss how the results can further be extended to additionally allow for heteroskedasticity and correlation in the error term.

If we only assume that

$$\mathbb{E}(\boldsymbol{u}) = \boldsymbol{0}, \qquad \|\mathbb{E}(\boldsymbol{u}\boldsymbol{u}')\|_2 \leq \overline{\sigma}^2, \tag{4.6}$$

we have the following result.

Theorem 5 (Generalized first-order approximation)**.** *Suppose that* (2.1) *is weakened by imposing only* (4.6)*. Let* $\mathcal{G}$ *be connected. Then*

$$\sqrt{d_i}\,(\widehat{\alpha}_i - \alpha_i) = \frac{1}{\sqrt{d_i}}\sum_{j\in[i]} u_{ij} + \epsilon_i\,,$$

where $\mathbb{E}(\epsilon_i) = 0$ *and* $\mathbb{E}(\epsilon_i^2) \leq \overline{\sigma}^2/(\lambda_2 h_i)$*.*

It follows that

$$\widehat{\alpha}_i \overset{a}{\sim} N(\alpha_i, \omega_i^2/d_i)$$

if $d_i^{-1/2} \sum_{j \in [i]} u_{ij} \xrightarrow{d} N(0, \omega_i^2)$ for finite ω_i^2, provided $\mathbb{E}(\epsilon_i^2) = o(1)$, which follows from $\lambda_2 h_i \to \infty$ and $\overline{\sigma}^2 < \infty$. The key asymptotic condition that $\lambda_2 h_i \to \infty$ thus remains as before. All discussion and examples from the previous section are thus also applicable to the more general situation of heteroscedastic and weakly correlated errors, now with ω_i^2 featuring in the asymptotic variance.

5 Weighted graphs

So far we have considered simple graphs. Our variance bounds generalize to weighted graphs. Let $\mathcal{G}$ be an undirected weighted graph with associated (weighted) adjacency matrix $\boldsymbol{A}$. A simple example is a multigraph, which differs from a simple graph in that multiple edges may exist between vertices. In this case, $(\boldsymbol{A})_{ij}$ equals the number of edges between i and j. More generally, $\boldsymbol{A}$ is symmetric, has diagonal entries equal to zero, and has off-diagonal entries that are non-negative.

Our variance bounds generalize to situations where an estimator $\check{\boldsymbol{\alpha}}$, constructed from $\mathcal{G}$, has variance $\boldsymbol{L}^\dagger$ for

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A},$$

where, as before, $\boldsymbol{D}$ is a diagonal (weighted) degree matrix with entries $d_i = \sum_{j=1}^n (\boldsymbol{A})_{ij}$. A symmetric matrix $\boldsymbol{L}$ is such a Laplacian matrix if and only if

(i) All off-diagonal elements of $\boldsymbol{L}$ are negative;

(ii) All column sums of $\boldsymbol{L}$ are equal to zero;

(iii) $\operatorname{rank}(\boldsymbol{L}) = n - 1$.

The variance bounds in Theorems 2 and 3 continue to apply to $\operatorname{var}(\check{\boldsymbol{\alpha}}) = \boldsymbol{L}^\dagger$ after setting $\sigma = 1$ and redefining the harmonic mean of the neighboring degrees to its weighted variant

$$h_i := \left(\frac{1}{d_i} \sum_{j \in V} (\boldsymbol{A})_{ij} \, d_j^{-1} \right)^{-1}.$$

The proofs of our theorems fully cover the weighted-graph case. We now give some examples of weighted graphs and corresponding estimators.

Example 3 (Panel data and weighted least squares)**.** Consider a situation where $(i,j) \in E$ interact on $m_{ij} \geq 1$ occasions and errors are heteroskedastic across meetings. Using obvious notation, the weighted least-squares estimator of $\boldsymbol{\alpha}$ equals

$$\check{\boldsymbol{\alpha}} := \arg \min_{\boldsymbol{a} \in \{\boldsymbol{a} \in \mathbb{R}^n : \boldsymbol{a}'\boldsymbol{\iota}_n = 0\}} \sum_{(i,j)\in E} \sum_{k=1}^{m_{ij}} \left(\frac{y_{ijk} - (a_i - a_j)}{\sigma_k} \right)^2 .$$

Let $\boldsymbol{A}$ be the weighted adjacency matrix with entries

$$(\boldsymbol{A})_{ij} := \begin{cases} \sum_{k=1}^{m_{ij}} \sigma_k^{-2} & \text{if } (i,j) \in E \text{ or } (j,i) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

and let $\boldsymbol{L}$ be the associated Laplacian matrix. Then Theorem 1 can be suitably extended to yield $\check{\boldsymbol{\alpha}} \sim N(\boldsymbol{\alpha}, \boldsymbol{L}^\dagger)$. □

Example 1 (cont'd) (Bipartite graph)**.** As before, consider a bipartite graph $\mathcal{G}$ where V is partitioned as $V_1 \cup V_2$ and edges are formed between V_1 and V_2 but not within these sets. Let $n_1 := |V_1|$ and $n_2 := |V_2|$. The Laplacian is

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A} = \begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_2 \end{pmatrix} - \begin{pmatrix} \boldsymbol{0} & \boldsymbol{C} \\ \boldsymbol{C}' & \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are $n_1 \times n_1$ and $n_2 \times n_2$ diagonal degree matrices and $\boldsymbol{C}$ is the $n_1 \times n_2$ upper-right block of the adjacency matrix of the graph. Decompose $\boldsymbol{\alpha}$ accordingly as $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')'$. The corresponding estimator $\widehat{\boldsymbol{\alpha}}$ is defined in (2.4) for the case of a simple graph, but the following construction works for any estimator that satisfies $\text{var}(\widehat{\boldsymbol{\alpha}}) = \sigma^2 \boldsymbol{L}^\dagger$, with $\boldsymbol{L}$ being the Laplacian matrix of a simple, weighted or multigraph, as described above (we may simply set $\sigma = 1$). We also define

$$\check{\boldsymbol{\alpha}}_2 := \widehat{\boldsymbol{\alpha}}_2 - \overline{\widehat{\boldsymbol{\alpha}}}_2 , \qquad \overline{\widehat{\boldsymbol{\alpha}}}_2 := \frac{1}{n_2} \sum_{i \in V_2} \widehat{\alpha}_i ,$$

corresponding to the natural normalization $\boldsymbol{\iota}_{n_2}'\check{\boldsymbol{\alpha}}_2 = 0$. By the block matrix inversion formula we find

$$\operatorname{var}(\check{\boldsymbol{\alpha}}_2) = \check{\boldsymbol{L}}^{\dagger}, \qquad \check{\boldsymbol{L}} := \sigma^{-2}\left(\boldsymbol{D}_2 - \boldsymbol{C}'\boldsymbol{D}_1^{-1}\boldsymbol{C}\right). \tag{5.7}$$

This is the variance formula after profiling-out all the parameters corresponding to vertices in V_1. It can be verified that $\check{\boldsymbol{L}}$ satisfies the conditions (i)–(iii) above. The adjacency matrix of the corresponding graph, say $\check{\mathcal{G}}$, that involves only the vertices in V_2 is given by the off-diagonal part of $\sigma^{-2}\,\boldsymbol{C}'\boldsymbol{D}_1^{-1}\boldsymbol{C}$. Thus, even when starting with a simple bipartite graph $\mathcal{G}$ we naturally obtain a weighted graph $\check{\mathcal{G}}$ when profiling out some of the parameters. Moreover, two vertices in $\check{\mathcal{G}}$ are connected if and only if they are path-two neighbors in the original graph $\mathcal{G}$. □

To illustrate, consider the simple wage regression of Abowd, Kramarz and Margolis (1999) with panel data where, now, the log wage of worker i in firm j at time t is equal to

$$y_{ijt} = \mu_i + \eta_j + u_{ijt}, \qquad t = 1, \ldots, m_{ij}.$$

To maintain focus, assume that the u_{ijt} are i.i.d. Then, with $\boldsymbol{\alpha} = (\boldsymbol{\mu}', -\boldsymbol{\eta}')'$ as discussed before, the pooled (ordinary) least squares estimator satisfies

$$\operatorname{var}(\widehat{\boldsymbol{\alpha}}) = \sigma^2\,\boldsymbol{L}^{\dagger},$$

where $\boldsymbol{L}$ is the Laplacian associated with the adjacency matrix

$$(\boldsymbol{A})_{ij} = (\boldsymbol{A})_{ji} = \begin{cases} m_{ij} & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

This illustration is interesting because, here, the μ_i cannot be estimated precisely, because the number of observations per worker is usually small. It therefore makes sense to focus on the η_j, that is, on estimating the firm effects. Profiling-out $\boldsymbol{\mu}$ and letting

$$\check{\boldsymbol{\eta}} := \widehat{\boldsymbol{\eta}} - \overline{\widehat{\boldsymbol{\eta}}}\,, \qquad \overline{\widehat{\boldsymbol{\eta}}} := \frac{1}{n_2}\sum_{i \in V_2} \widehat{\eta}_i\,,$$

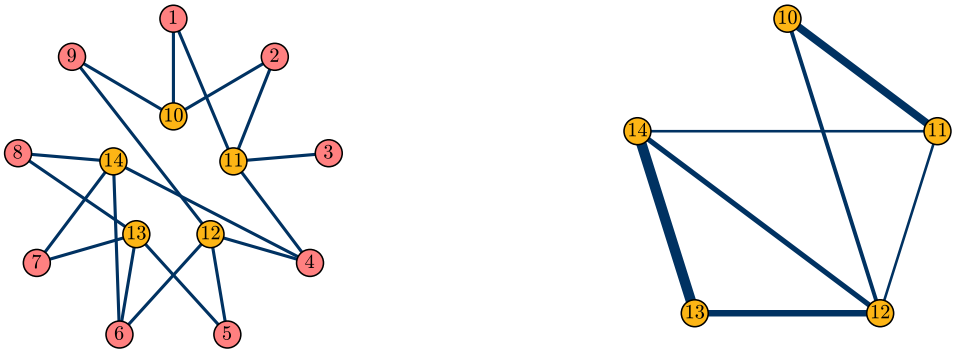

Figure 1: A simple bipartite graph $\mathcal{G}$ (left) with links between V_1 (red vertices) and V_2 (yellow vertices), and the induced weighted graph $\check{\mathcal{G}}$ (right) on V_2 alone resulting from profiling out the parameters associated with V_1.

where $n_2 := |V_2|$ is the number of firms, we find as an application of (5.7) that

$$\mathrm{var}(\check{\boldsymbol{\eta}}) = \check{\boldsymbol{L}}^{\dagger},$$

where $\check{\boldsymbol{L}}$ is the $n_2 \times n_2$ Laplacian matrix associated with the weighted $n_2 \times n_2$ adjacency matrix

$$(\check{\boldsymbol{A}})_{jk} := \begin{cases} \sigma^{-2} \displaystyle\sum_{i \in [j] \cap [k]} \frac{m_{ij}\, m_{ik}}{d_i} & \text{for } j \neq k, \\ 0 & \text{for } j = k, \end{cases}$$

where $d_i = \sum_{j \in V} m_{ij}$ is the degree of $i \in V_1$, that is, the total number of observations for that worker, and $[j] \cap [k]$ is the set of all workers for which wages are observed both in firm j and in firm k. In this example the vertex set of of the weighted graph $\check{\mathcal{G}}$ are the firms. Two firms are connected by an edge if there is at least one worker who has worked in both firms. The weight $(\check{\boldsymbol{A}})_{jk}$ of the edge is larger the more workers there are connecting firms j and k, and the longer these workers have worked in both firms. Figure 1 provides an illustration of a simple bipartite graph (with all $m_{ij} = 1$) for workers (red vertices) and firms (yellow vertices), given in the left plot, and the induced weighted graph featuring only firms, given in the right plot. The thickness of the edge between (j, k) in the plot of $\check{\mathcal{G}}$ reflects the magnitude of the weight $(\check{\boldsymbol{A}})_{jk}$.

References

Aaronson, D. L., L. Barrow, and W. Sander (2007). Teacher and student achievement in the Chicago public high schools. *Journal of Labor Economics 25*, 95–135.

Abowd, J., R. Creecy, and F. Kramarz (2002). Computing person and firm effects using linked longitudinal employer-employee data. U.S. Census Technical Paper TP-2002-06.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*, 251–333.

Andrews, M. J., L. Gill, T. Schank, and R. Upward (2008). High wage workers and low wage firms: Negative assortative matching or limited mobility bias. *Journal of the Royal Statistical Society, Series A 171*, 673–697.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika 39*, 324–325.

Bramboullé, Y., R. Kranton, and M. D'Amours (2014). Strategic interaction and networks. *American Economic Review 104*, 898–930.

Card, D., J. Heining, and P. Kline (2013). Workplace heterogeneity and the rise of West German wage inequality. *Quarterly Journal of Economics 128*, 967–1015.

Chung, F. R. K. (1997). *Spectral Graph Theory*. Volume 92 of CBMS Regional Conference Series in Mathematics, American Mathematical Society.

Erdős, P. and A. Rényi (1959). On random graphs. *Publicationes Mathematicae 6*, 290–297.

Finkelstein, A., M. Gentzkow, and H. Williams (2016). Sources of geographic variation in health care: Evidence from patient migration. *Quarterly Journal of Economics 131*, 1681–1726.

Hoffman, C., M. Kahle, and E. Paquette (2013). Spectral gaps of random graphs and applications to random topology. Mimeo.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review 94*, 247–252.

Schiebinger, G., M. J. Wainwright, and B. Yu (2015). The geometry of kernelized spectral clustering. *Annals of Statistics 43*, 819–846.

Verdier, V. (2016). Estimation and inference for linear models with two-way unobserved heterogeneity and sparsely matched data. Mimeo.