A Model for Path Data

Dean Knox*

This draft: August 30, 2016

Abstract

Path data describes the steps that an actor takes to get from point A to B. It offers researchers the opportunity to test theories about network navigation, including in social and geographic networks. For example, path data can show whether individuals avoid out-group neighborhoods in their daily walking routes, resulting in societal inefficiencies and reducing inter-group contact. This data can also reveal how voters search social networks for political information, which may distort the information they ultimately receive. However, the sequential decision-making process in path data violates the underlying assumptions of existing models, which assume some form of conditional independence between observations. I propose a new random-path model (RPM) that explicitly captures this pathwise dependence, develop an estimation procedure, and demonstrate its properties. The RPM builds on a random-walk model, incorporating a realistic but difficult-to-analyze constraint to account for the fact that actors are purposefully navigating toward a destination. I validate the model in an analysis of the U.S. Interstate Highway planning process, where existing approaches fail to recover a known qualitative benchmark. Finally, the RPM is used to test two competing explanations of Baghdad's recent segregation. Using smartphone-based behavioral data from Sunni and Shia participants in a field activity, I show that a *need-based* model of residential sorting—when families flee mixed neighborhoods to avoid political violence—is insufficient to explain participants' walking routes alone. Instead, their choices reveal that conflict has also created significant *taste-based* aversion to out-groups in a city once known for its cosmopolitanism. These results suggest that societal preferences have shifted in a way that makes Baghdad's eventual re-integration unlikely.

*Ph.D. Candidate, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: dcknox@mit.edu.

1 Introduction

Goals are rarely accomplished in one fell swoop. Instead, actors work one step at a time, making a series of smaller decisions that ultimately lead to the intended destination. This sequential process forms a *path*. As a general phenomenon, paths are ubiquitous, and social scientists have developed a variety of theories about them. For example, sociologists are interested in how people search their social networks through a chain of intermediaries and referrals (Milgram, 1967; Killworth and Bernard, 1978; Dodds, Muhamad and Watts, 2003), and development economists seek to evaluate the impact of transportation infrastructure, such as highway and railroad routes, on long-term growth (Fogel, 1962, 1964; Aschauer, 1989; Banerjee, Duflo and Qian, 2012). In these theories, paths feature as both dependent and independent variables—just like any other form of data. Yet paths are rarely measured and studied as such when evaluating path-related theories. In this, they differ from other data types, such as event counts, where a broad consensus has emerged that theories should be tested by collecting appropriate forms of data and using appropriate models to analyze them (King, 1989). The gap is largely due to the absence of statistical models for path data.

In this paper, I propose and demonstrate the properties of a new model for path outcomes. Paths are represented as movement on a network: A forward-looking decision-maker chooses from a limited set of options, or neighboring nodes, each opening a new and different set of possible next steps. For example, when building a road to connect two cities, a planner must go through an inner suburb of the first, then an adjacent outer suburb, and so on until eventually arriving at the destination. The random-path model (RPM) allows researchers to learn about preferences based on this movement. The intuition underlying the model is simple. If roads commonly go out of their way to avoid mountainous regions, then the "cost" of elevation is larger than deviating from the shortest route.

The RPM builds on the random-walk model, which assumes that each decision is independent.

Random walks are powerful and well-studied models that can explain how long-term patterns emerge from a series of small decisions. They have found application across a wide range of fields, including the random collisions of atoms, foraging patterns of animals, and stock-market fluctuations. In the random-walk model, a walker myopically takes steps based on their short-term attractiveness, until eventually reaching the goal by sheer luck. As an example, consider an individual navigating between diagonal corners of a 5×5 street grid. Under this model, the walker has a 99.2% chance of wandering back to a previously visited intersection at some point before arriving at the destination. While the random walk might be appropriate model of human decision-making in some circumstances (as a case in point, this particular problem is often called the "drunkard's walk"), I argue that actors generally plan ahead and work more efficiently.

To better model purposeful decision-making, the RPM adds a conditioning stage before starting, in which all candidate routes that contain repetitive loops are discarded. It can be shown that this is equivalent to a forward- and backward-looking walker that navigates toward the goal and avoids previously visited areas. In the same street-grid example, under the RPM, the walker pushes onward to the destination rather than wandering aimlessly in circles. Moreover, under this simple scenario, a walker following the RPM takes a shortest path—walking eight blocks, e.g. along the diagonal—nearly three-quarters of the time. (Additional covariates can be incorporated to account for the walker's sense of direction or familiarity with the area.) A model that incorporates some form of long-term planning is better-suited for most applications in political science.

However, this additional constraint poses a challenge in that it makes the resulting probability mass function intractable. I demonstrate that despite this challenge, the RPM can still be estimated. To this end, I develop and assess numerical algorithms for sampling random paths, evaluating a simulated RPM likelihood function, and efficiently implementing Metropolis-Hastings sampling from the posterior distribution. A permutation-based extension of the model can be used to estimate the causal effects of path-assigned treatments.

In the remainder of this paper, I first discuss examples of path data in political science. Section 3 formally defines the model, outlines the estimation procedure, and contrasts the random path model with existing approaches. Section 4 validates the RPM with a study of the U.S. Interstate Highway System, where official priorities are known from detailed qualitative planning documents. I show that RPM estimates correspond closely to this benchmark, whereas existing spatial models produce substantively and statistically differing results.

Finally, section 5 provides a motivating empirical application. In Baghdad, a long history of Sunni–Shia coexistence and integration was overturned by a wave of ethnic cleansing in 2006–2007. The long-run effects of this conflict depend on whether recent segregation was only driven by need-based sorting—to avoid violence (Morrison, 1993)—or whether it also led to taste-based sorting (Schelling, 1969, 1971). While these drivers of out-group aversion often go hand-in-hand, they have very different implications for Baghdad's future development: If residents only fled mixed neighborhoods to avoid ethnic violence, then game-theoretic models (Young, 1998; Zhang, 2004*a*,*b*) predict gradual post-conflict reintegration. If conflict led to newfound taste-based aversion, however, ethnic attitudes can persist far beyond the end of conflict and make segregation difficult to escape.

To test these competing hypotheses, I analyze behavioral data from a Baghdad field study by Christia and Knox (n.d.). Participants' movement in a treasure-hunt-type activity reveals that taste-based aversion is a significant factor in participants' daily movement. These results suggest that recent conflict has shifted societal preferences in a way that makes reintegration unlikely. Section 6 concludes with limitations and areas for future work.

2 Literature Review

2.1 Path Data in Political Science

Political scientists have theorized about the causes and effects of paths in a variety of settings. In this section, I offer examples that include the geographic paths of connective infrastructure, the social paths traced by individuals as they search their social networks, and the aggregate-flow paths of people and goods. Other path-related theories appear in economics, urban planning, operations research, and engineering. While the measurement of path *data* is relatively new, it is growing increasingly common, offering new opportunities for research but also presenting new challenges for statistical analysis.

An illustrative example of a path in political science is the highway, which connects cities through a series of intermediate counties. Researchers are often interested in evaluating the role of various political factors in transportation spending, including institutions, pork-barrel spending, and ethnic patronage (e.g. Lee, 2000; Burgess et al., 2015). Other examples from connective infrastructure include the impact of patronage in electrical grid construction (Briggs, 2012) and governance on oil pipelines (Carmody, 2009). Researchers have increasingly recognized that standard models fail to account for the dependence between units that arises in these contexts. That is, whether a county is connected to the highway system depends not only on *whether* neighboring counties are also connected, as in standard spatial models, but also *which* neighbors are connected.[1] Given that highways are typically designed to connect major metropolitan areas, a rural legislator's success in securing transportation spending is perhaps as much about diverting the course of already-planned segments as it is fabricating entirely new projects. However, existing models do not permit principled testing of hypotheses about factors shaping the trajectories of paths. I argue that paths are a unique class of dependent variable and should be modeled

[1] For example, consider an east-west highway on a square grid. For county i to be connected, two conditions are required: (1) Its neighbors to the east and west must be connected, so that i can serve as the missing link; and (2) neighbors to the north and south *cannot* be connected—otherwise, the route would already be complete and i would be superfluous.

accordingly, much as event count outcomes are commonly modeled with Poisson regression.

Social scientists are not only interested in modeling paths as a dependent variable, but also in evaluating the effects of path-assigned treatments. In economics, a long-standing debate revolves around whether connective infrastructure leads to economic development (Fogel, 1962, 1964; Aschauer, 1989; Banerjee, Duflo and Qian, 2012; Casaburi, Glennerster and Suri, 2013). More recent work has also linked highway construction to popular support for the Nazi party (Voigtlaender and Voth, 2014) and urban–suburban political polarization (Nall, 2013, 2015); electrification to liberalizing attitudes in the Tennessee Valley (Caughey, 2012); and oil pipelines to local revenue sharing (Blair, 2016). This work has generally relied on context-specific information to address inferential challenges in these settings—for example, approximating an ideal experiment in which several alternative highways are proposed, but only some selected for construction. I discuss a way to take this intuition and generalize it, by modeling the path assignment process.

Beyond connective infrastructure, paths also appear widely in studies of social network search. As a concrete example, Habyarimana et al. (2007) showed that co-ethnic networks increase "findability" of strangers in their study of public goods provision. In their experiment, randomly selected Ugandan "runners" were given photographs of strangers, then asked to locate them within 24 hours—which they did with startling levels of success. The ability to locate and sanction free riders through social networks, often easier among co-ethnics, is in turn linked to public goods provision in diverse societies (Miguel and Gugerty, 2005; Eubank, 2016). Christia, Knox and Al-Rikabi (n.d.) examine related questions in the context of Iraq, showing that minority Sunnis adapt to a Shia-dominated society by developing more efficient network search strategies to access public services. Network search is also important in the spread of political information, where the search patterns of citizens actively seeking knowledge can distort the information they ultimately receive (Huckfeldt and Sprague, 1987, 1995). (The passive transmission of political information, like the spread of rumors, can also be thought of as a flow path over a network (Converse, 1962; Zaller, 1989).) Finally, the formation of buyer-seller ties in imperfect markets often involves searching a

social network for exchange partners (Kranton and Minehart, 2001).

Path outcomes are also of great interest in public policy, where they can represent aggregate flows of, e.g., refugee migration or smuggled drugs. Potential determinants of migration routes, such as welfare policy and border security, have been debated with increasing urgency since a sharp increase in European migration in 2015 (U.N. High Commissioner for Refugees, 2016). The spillover effects of policies that divert such flows is a common concern. For example, Dell (2015) considers a game-theoretic model of path-based spillover, in which crackdowns by Mexico's National Action Party (PAN) force drug traffickers to reroute through nearby municipalities, and shows that the model is consistent with rising drug arrests after PAN victories.

The random-path model allows researchers to learn about the preferences of an actor, such as the highway planners, information-seeking citizens, and drug traffickers discussed above, based on an observed path or collection of paths. It can test whether patronage is a significant factor in the trajectories of highways and electrical lines—that is, whether infrastructure routes deviate from an "optimal" route in order to visit certain areas—or whether drug traffickers tend to avoid states with harsher penalties, such as minimum-sentencing laws. Furthermore, the model allows researchers to simulate various quantities of interest in a statistically principled way: How many additional miles of highway were built because of patronage? If one state cracked down on trafficking, what volume of drug shipments would divert into its neighbors? Finally, in ongoing work, I show that RPM can be used as a model of treatment assignment to allowing inference about the causal effects of path-assigned treatments (Rubin, 1991). This approach also permits the study of spillover effects (Bowers, Fredrickson and Panagopoulos, 2013; Aronow and Samii, n.d.), such as whether highways lead to growth in nearby areas or whether they contribute to out-migration and decline, by comparing highway towns and nearby areas to places that were as likely to be connected.

2.2 Alternative Methods

Most analyses of paths, such as infrastructure, have used standard regression or matching methods that ignore spatial dependence entirely (Rephann and Isserman, 1994; Chandra and Thompson, 2000; Michaels, 2008; Donaldson, forthcoming). Others allow for correlation within a cluster of units, such as counties in a state, but neglect the fact that bordering counties on opposite sides of a state line are *also* highly dependent (Baum-Snow, 2007; Baum-Snow et al., n.d.). At best, researchers have employed spatial error or spatial autocorrelation models that assume dependence is (1) isotropic, so that each unit is positively correlated with its circular neighborhood;[2] (2) decaying at a constant rate with distance; and (3) stationary, so that the same correlation structure is constant the entire space (Cohen and Paul, 2004; Del Bo and Florio, 2012).

Spatial models are well-suited for analyzing a variety of phenomena, such as policy diffusion (Elkins and Simmons, 2005). However, they are not intended for the analysis of path data, which violate every one of the underlying assumptions outlined above. As a concrete example, consider the naturalistic simulation in appendix B, where a road curves around a mountain range to connect cities on opposite sides. Path data has strong positive dependence in some directions (for a town to be connected, the road must approach it from the front and back) and negative in others (if the road detours around one side, it will not pass through the town). Roads can easily exhibit long-range dependence—all towns on one side of the mountain are positively correlated with each other, and they are negatively correlated with towns on the opposite side—the road can only choose one side, and it connects many areas on that side simultaneously. Moreover, the distribution over possible roads is tighter in a mountain pass, where fewer viable routes exist.

Generally speaking, analyses that ignore pathwise dependence between observations lead to results that are as much of an "exercise in self-deception" as those that ignore clustering (Cornfield, 1978). Moreover, all of the models describe above essentially treat local dependence as a nuisance. Their output cannot be interpreted in terms of useful path-related quantities of interest—for

[2]Some analyses relax this assumption to allow for elliptical correlation structures.

example, the change in road length caused by the mountain described above.

In contrast to spatial models, the RPM is a model of network formation. It starts with the graph of all neighboring nodes, such as counties, and selects a subset of contiguous nodes and edges to connect a starting point to an endpoint. Other families of network models that have been used in political science include exponential random graph models (ERGMs) and latent-space models (Cranmer et al., 2016). Broadly speaking, existing network models deal with dyadic relationships while accounting for the contextual influence of local network structure. For example, ERGMs have proven valuable in the study of international relations for their ability to account for the way allies' relationships affect whether two nations go to war (Cranmer and Desmarais, 2011). The goal of the RPM differs from these models in that it models a purposeful attempt to connect two nodes; it is influenced by network structure over a much longer range, is subject to more severe constraints, and generally addresses a different class of questions.

3 Model

In this section, I briefly discuss two interpretations of random walks, which are closely related to the random-path model. Walks are introduced as a sequence of dependent random steps. This view is then shown to be mathematically equivalent to an alternative view in which entire walks are drawn, all at once, from a discrete set of sequences. I exploit this equivalence to conveniently express random-path models in the second view, then discuss the implied relationship between random-walk models and RPMs in the first. I then outline the computational challenges in estimating RPMs and outline a procedure to recover the posterior distribution of the random-path parameters, given a set of observed paths. The method is placed in the context of the simulated likelihood method and a rapidly growing literature on approximate Bayesian computation.

3.1 Random Walks: A Review

Define a weighted, possibly directed graph G as a set of nodes (vertices) denoted $V \in \{1, \cdots, N\}$, such as counties, and a row-stochastic edge-weight matrix $E = [\varepsilon_{i,j}] = [\boldsymbol{\varepsilon}_{1,*}^\top, \cdots, \boldsymbol{\varepsilon}_{N,*}^\top]^\top$. For a walker at i, $\varepsilon_{i,j}$ represents the probability that the walker's next step is to j; it takes on positive values for adjacent j—those in i's neighborhood, $\mathcal{N}_i$, which is the "choice set" for a walker at i—and zero otherwise. Self-links, or $\varepsilon_{i,i}$, are set to zero by convention.

A random walk, $\Gamma \equiv (v_0, \cdots, v_K)$, is defined by a starting node v_0, the transition distributions $v_t \sim \text{Categorical}(\boldsymbol{\varepsilon}_{t-1,*})$, and a stopping rule.[3] For illustrative purposes, I assume that walks stop upon reaching a single predesignated terminus, v_K. The observed path is denoted $\gamma = (\gamma_0, \cdots, \gamma_k)$, and the specified conditions require that $v_0 = \gamma_0$ and $v_K = \gamma_k$. Note that the number of steps in the walk, K, is also a random variable, with realization k. (Alternative stopping rules, such as after a fixed number of steps or when any of a set of nodes are found, may be more appropriate in other applications, and the proposed distribution is easily adapted for these cases.)

The random walk is analogous to the negative binomial distribution in that it can be thought of as either a sequence of dependent categorical random variables, as presented above, or a probability distribution over an infinite discrete set whose elements are sequences of varying length. In either case, given fixed endpoints, the probability of a particular realization is

$$\Pr(\Gamma = \gamma \mid v_0 = \gamma_0, v_K = \gamma_k) = \prod_{t=0}^{k-1} \varepsilon_{\gamma_t, \gamma_{t+1}}$$

It is straightforward to model step probabilities, $\varepsilon_{i,j}$, as a function of M covariates. Let $\boldsymbol{X}$ be a $N \times N \times (M+1)$ tensor where the m-th slice is a matrix of dyadic covariates, such as distance. $\beta = [\beta_0, \beta_1, \cdots, \beta_M]^\top$ is a vector of coefficients, and $\boldsymbol{X}_m \beta^m = \left[\sum_m \beta_m X_{*,*,m}\right]$ is a $n \times n$ matrix

[3] This defines a walk in terms of a node sequence, which leaves the intervening edges $v_t v_{t+1}$ implicit. An equivalent definition is that a walk is a subgraph of G, $G_\Gamma = (V_\Gamma, E_\Gamma)$, in which $V_\Gamma \subseteq V$ and E_Γ is a sequence of edges, $(v_0 v_1, \cdots, v_{K-1} v_K)$, in which all elements satisfy $\varepsilon_{v_t, v_{t+1}} > 0$.

of linear predictors. Assume edge weights can be written as

$$\varepsilon_{i,j} = \frac{\exp\left(\left[\boldsymbol{X}_m \beta^m\right]_{i,j}\right)}{\sum_{j'} \exp\left(\left[\boldsymbol{X}_m \beta^m\right]_{i,j'}\right)},$$

so that rows of E are the multinomial logistic transformation, or softmax, of rows of $\boldsymbol{X}_m \beta^m$. Fix β_0 at $-\infty$ and let $X_{*,*,0} = [\mathbf{1}(j \notin \mathcal{N}_i)]$, so that $\varepsilon_{i,j} = 0$ for $j \notin \mathcal{N}_i$, and the probability mass function (PMF) is

$$f_{\text{walk}}(\gamma \mid \boldsymbol{X}, \beta) \;\equiv\; \Pr(\Gamma = \gamma \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta) \;=\; \prod_{t=0}^{k-1} \frac{\exp\left(\left[\boldsymbol{X}_m \beta^m\right]_{\gamma_t, \gamma_{t+1}}\right)}{\sum_{j'} \exp\left(\left[\boldsymbol{X}_m \beta^m\right]_{\gamma_t, j'}\right)}$$

The random-walk model is well-understood and has been used in the transportation literature. Fosgerau, Frejinger and Karlstrom (2013) observe that these models allow loops, including infinite loops, although they report that these are rare in their particular case.

3.2 Random Path Distribution as Conditional Random Walk

The random walk, while analytically tractable, is a poor model for many social phenomena because it assumes that each step is independent. Unlike sequential decision-makers in political science, such as highway planners, random walkers are neither forward- or backward-looking. Under typical conditions, they are likely to revisit many nodes—as I describe in section 1, a random walker crossing a $N \times N$ street grid will go in circles with near certainty for $N \geq 5$. Moreover, this problem cannot be fixed by incorporating covariates into the step probabilities (such as distance or direction), because loops are a property of the entire sequence rather than any particular step.

This paper proposes an alternative, the conditional random-walk distribution $(\Gamma \mid \Gamma \in \mathcal{P})$, where $\mathcal{P}$ is the set of all possible paths from γ_0 to γ_k—i.e., all walks from start to terminus that contain no loops. Formally, $\mathcal{P} \equiv \{\psi : \Omega_\Gamma, |\{\psi\}| = |\psi|\}$, where Ω_Γ is the sample space of Γ and the latter condition specifies that all nodes in path ψ are unique. Thus, $\mathcal{P}$ excludes all walks

that return to a previously visited node. Given that the observed walk γ is a path, so that it automatically satisfies $\gamma \in \mathcal{P}$, the random-path PMF is found by renormalizing:

$$\begin{aligned} f_{\text{path}}(\gamma \mid \boldsymbol{X}, \beta) &\equiv f_{\text{walk}}(\gamma \mid \boldsymbol{X}, \beta, \Gamma \in \mathcal{P}) \\ &= \frac{\Pr(\Gamma = \gamma, \Gamma \in \mathcal{P} \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)}{\Pr\left(\Gamma \in \mathcal{P} \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta\right)} \\ &= \frac{\Pr(\Gamma = \gamma \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)}{\Pr\left(\Gamma \in \mathcal{P} \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta\right)} \\ &= \frac{\prod_{t=0}^{k-1} \frac{\exp\left([\boldsymbol{X}_m \beta^m]_{\gamma_t, \gamma_{t+1}}\right)}{\sum_{j'} \exp\left([\boldsymbol{X}_m \beta^m]_{\gamma_t, j'}\right)}}{\sum_{\psi \in \mathcal{P}} \prod_{t=0}^{|\psi|-1} \frac{\exp\left([\boldsymbol{X}_m \beta^m]_{\psi_t, \psi_{t+1}}\right)}{\sum_{j'} \exp\left([\boldsymbol{X}_m \beta^m]_{\psi_t, j'}\right)}}. \end{aligned} \tag{1}$$

This random-path distribution has properties that make it well-suited for modeling common decision-making processes, such as the political science applications discussed in section 2.1. Recall that in the view of the random walks as a sequence of random variables, at each step, the walker is only "backward-looking" insofar as the previous step determines the current options. That is, in a random walk, $\Gamma_t \not\perp\!\!\!\perp \Gamma_{t-1}$, but $(\Gamma_t \mid \Gamma_{t-1}) \perp\!\!\!\perp \Gamma_{t-2}$. In the same view of random paths, the walker is "fully" backward-looking in that it will tend to avoid the vicinity of all previously visited nodes. The walker is also forward-looking in that it tends to avoid traps and other local optima with foresight, anticipatorily moving in directions that will take it to the destination faster.

3.3 Estimation

Equation 1 suggests a likelihood-based approach for inference on the random-path model. I begin by briefly discussing an algorithm for exactly calculating this likelihood. Because this approach becomes intractable for moderately sized or dense graphs, I then develop an simulation-based approximation that converges to the exact method as the number of simulations tends to infinity. Finally, I briefly discuss computational issues for Bayesian inference on RPM models.

3.3.1 Simulated RPM Likelihood

A natural approach for inference on random paths is to use the likelihood $\mathcal{L}_{\text{path}}(\beta \mid \boldsymbol{X}, \gamma) \equiv f_{\text{path}}(\gamma \mid \boldsymbol{X}, \beta)$, where the full expression for the right-hand side is given in equation 1. The chief difficulty in doing so is that the denominator of equation 1 varies with β and involves summing over the (typically large) set of possible paths between the observed start- and endpoints, γ_0 and γ_k. For example, the maximum likelihood estimate of β are the parameters that maximize the ratio of (a) the unconditional (random-walk) probability of the observed path to (b) the totaled random-walk probabilities of every other path that could have been drawn. In appendix A.1, I describe an exact method for doing so. This procedure uses a recursive search to explicitly enumerate every possible path, then sums the random-walk probabilities of mutually exclusive paths. In principle, the Fisher information matrix can also be derived in this way, with confidence intervals calculated using the delta method.

However, in practice, explicitly enumerating and operating on all possible paths is computationally infeasible, even for moderately sized or dense graphs. For example, in complete graphs, where every node is connected to every other, the number of possible paths is given by $\sum_{k=0}^{N-2} \frac{(N-2)!}{k!}$. Even in a ten-node complete graph, 109,601 paths are possible. Building on the intuition behind the exact approach, I develop algorithm 1 to approximate the likelihood function to arbitrary precision. Algorithm 1 is based on a common numerical approach for summations over hard-to-enumerate domains, Monte Carlo integration. The approach developed here is analogous to the following procedure for approximating the integral $\int_a^b f(x)\ dx$: Randomly sample points on the uniform $[a, b]$ distribution, evaluate $f(x)$ at each point, average the results, and multiply by the size of the sampling space $(b - a)$.

To apply this to the RPM case, let Ψ be the uniform distribution over the set of all possible paths $\mathcal{P}$. The following is a directly analogous approach for estimating the denominator of equation 1, which is equivalent to $\sum_{\psi \in \mathcal{P}} f_{\text{walk}}(\psi)$. Repeatedly draw $\psi \sim \Psi$, evaluate the random-walk probability for each sampled element, and average across draws to estimate

$\mathbb{E}_\Psi[f_{\text{walk}}(\Psi)] = \frac{1}{|\mathcal{P}|}\sum_{\psi\in\mathcal{P}} f_{\text{walk}}(\psi)$. Then, multiply by the total number of paths $|\mathcal{P}|$ to find an estimate of the denominator. Because the numerator is calculated exactly, the simulated likelihood (Lee, 1992) inherits the desirable property of converging to the exact likelihood in appendix A.1 (up to a multiplicative constant) as the number of or simulations tends to infinity. This can be seen by noting that depth-first search finds every path exactly once, and the number of times that the uniform distribution draws each path converges to $\frac{S}{|\mathcal{P}|}$ as S grows large.

There are two complications in this procedure. The lesser complication is that we need to know the value of $|\mathcal{P}|$ to correctly normalize. This is a #P-hard problem (Valiant, 1979), meaning that it can only be solved by listing every possible path and then counting them—precisely the computational issue that we were trying to sidestep in the first place.[4] Fortunately, $|\mathcal{P}|$ does not involve the RPM parameters, β, and so it can be absorbed into the normalizing constant of the RPM likelihood function. The RPM likelihood is then given by

$$\begin{aligned}
\mathcal{L}_{\text{path}}(\beta \mid \boldsymbol{X}, \gamma) &= \frac{\prod_{t=0}^{k-1} \frac{\exp\left([\boldsymbol{X}_m\beta^m]_{\gamma_t,\gamma_{t+1}}\right)}{\sum_{j'}\exp\left([\boldsymbol{X}_m\beta^m]_{\gamma_t,j'}\right)}}{\sum_{\psi\in\mathcal{P}}\prod_{t=0}^{|\psi|-1}\frac{\exp\left([\boldsymbol{X}_m\beta^m]_{\psi_t,\psi_{t+1}}\right)}{\sum_{j'}\exp\left([\boldsymbol{X}_m\beta^m]_{\psi_t,j'}\right)}} \\
&= \frac{\prod_{t=0}^{k-1} \frac{\exp\left([\boldsymbol{X}_m\beta^m]_{\gamma_t,\gamma_{t+1}}\right)}{\sum_{j'}\exp\left([\boldsymbol{X}_m\beta^m]_{\gamma_t,j'}\right)}}{|\mathcal{P}|\cdot\mathbb{E}_\Psi\left[\prod_{t=0}^{|\Psi|-1}\frac{\exp\left([\boldsymbol{X}_m\beta^m]_{\Psi_t,\Psi_{t+1}}\right)}{\sum_{j'}\exp\left([\boldsymbol{X}_m\beta^m]_{\Psi_t,j'}\right)}\right]} \\
&\propto \frac{\prod_{t=0}^{k-1} \frac{\exp\left([\boldsymbol{X}_m\beta^m]_{\gamma_t,\gamma_{t+1}}\right)}{\sum_{j'}\exp\left([\boldsymbol{X}_m\beta^m]_{\gamma_t,j'}\right)}}{\mathbb{E}_\Psi[\Pr(\Gamma=\Psi \mid v_0=\gamma_0, v_K=\gamma_k, \boldsymbol{X}, \beta)]}. \qquad (2)
\end{aligned}$$

After eliminating this constant, equation 2 could in principle be approximated by Monte-Carlo sampling from Ψ as described above. This brings us to the second complication. Unfortunately, Ψ cannot be sampled—there is no known algorithm for uniformly sampling paths. To deal with this, I adapt a non-uniform distribution for importance sampling on $\mathcal{P}$, the loop-erased random walk (LERW). [5] The LERW begins with a pure random walk, then retraces its steps and removes

[4] Roberts and Kroese (2007) explore an approximation for large graphs, but its accuracy is unknown.

[5] An alternative approach for sampling non-uniformly from $\mathcal{P}$ is the self-avoiding walk (SAW). The SAW is a

loops that return to previously visited nodes (the procedure is given as part of algorithm 1). Wilson (1996) proved that the spanning tree—i.e., a subgraph that connects all nodes, such as a maze, that contains no cycles—produced by iteratively combining LERWs will be a uniform draw from the set of all spanning trees. In proposition 1, I use this property to construct an importance-sampling scheme on $\mathcal{P}$.

Proposition 1 (Simulated RPM Likelihood)**.** *Define the unweighted version of G, $\tilde{G}$, and let L be a path-valued random variable, with distribution $f_{\text{LERW}}(\psi : \tilde{G}, v_0, v_K)$, that can be sampled by the loop-erased random walk on $\tilde{G}$ from v_0 to v_K. The denominator of equation 2 can be rewritten*

$$\begin{aligned}&\mathbb{E}_\Psi[\Pr(\Gamma = \Psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)] \\ &\quad = \sum_{\psi \in \mathcal{P}} \Pr(\Gamma = \psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)\ f_{\text{LERW}}(\psi :\ \tilde{G}, v_0, v_K)\ w(\psi),\end{aligned}$$

and its importance-sampling estimate is

$$\begin{aligned}&\hat{\mathbb{E}}_\Psi[\Pr(\Gamma = \Psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)] \\ &\quad = \sum_{s=1}^{S} \Pr(\Gamma = L_s \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)\ w(L_s),\end{aligned}$$

where S is the number of importance-sampling draws. The adjustment factor $w(\psi) \propto \frac{1}{\det\ L_{(-i,-j)}(\tilde{G}/\psi)}$ is the ratio between the target uniform distribution and the LERW distribution. The above holds for any i and j in V. The term $\tilde{G}/\psi$ is the iterated edge contraction of $\tilde{G}$ along all edges in path ψ, $L(\cdot)$ is the Laplacian matrix of a graph, and $M_{(-i,-j)}$ is the (i, j) minor of a square matrix M.

A proof is given in Appendix A.2. Briefly, Wilson (1996) implies that the probability that a LERW draws a particular path, ψ, will be proportional to the number of spanning trees that

random walk that sets transition probabilities to zero for previously visited nodes, as in *Snake* (Gremlin, 1976). Properties of the SAW are largely unknown, so this approach is not considered. As part of their algorithm, Roberts and Kroese (2007) attempt to estimate and correct for the bias of SAWs toward shorter paths by simulation. They employ an ad-hoc method that increases the probability of long paths by down-weighting transition probabilities to the final node (i.e., avoiding termination) based on the number of steps that have been taken. However, their approach under-samples nodes that are distant from the target, convergence rates of the various correction factors are unknown, and the resulting distribution is poorly understood.

include ψ as a subgraph. Proposition 1 uses the deletion-contraction recurrence to exclude trees that cannot arise under ψ, then applies Kirchoff's matrix tree theorem to the contraction to count the number of such spanning trees. Proposition 1 immediately suggests a simulated-likelihood analogue of equation 2. This simulated likelihood forms the basis for statistical inference.

3.3.2 Discussion

The methodological challenges that arise in the RPM are closely related to those in fixed-effects logistic regression model and related models. In the fixed-effect logit setting, conditioning on a sufficient statistic induces a combinatorics problem in the denominator of the conditional likelihood. This problem, which also arises in censored survival data, is commonly addressed with various analytical approximations (Breslow, 1974; Efron, 1977). In the RPM setting, the no-loop conditioning leads to a similar problem, but the combinatorics in the denominator are sufficiently complex that no closed-form approximations are known.

The simulated likelihood approach taken here is a response to this problem. It uses Monte Carlo simulations to integrate over the outcome space, and thus approximate the denominator, by computational rather than analytical methods. The procedure differs from typical applications of the simulated likelihood method, which use Monte Carlo integration to marginalize nuisance variables such as random coefficients (Bhat, 2001). A related approach to outcome-space integration is used in approximate Bayesian computation, when exact calculation of the likelihood is not practical or possible (for recent developments, see Marin et al., 2012).

This simulated likelihood is computationally intensive for two reasons. First, for each of the S sampled paths is drawn, an expensive matrix determinant must be calculated to find the adjustment factor. Second, each time the simulated likelihood is evaluated at some point in the parameter space, algorithm 1 loops over and operates on all S paths. (In the applications explored here, values of S are on the order of 10^6 paths).

Neither the simulated likelihood nor the true likelihood function to which it converges are

Data:
starting node γ_0, teminus γ_k, covariates $\boldsymbol{X}$, parameters β
unweighted graph $\tilde{G}$, number of simulations S

Result:
$\psi_1, \cdots, \psi_s \in \mathcal{P}$
$w_1, \cdots, w_s$, inverse importance weights
$\hat{\mathbb{E}}_\Psi[\Pr(\Gamma = \Psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)]$,
estimated denominator of random-path likelihood function (equation 2)

Algorithm `ApproxPrPath`$(\gamma_0,\ \gamma_k,\ \boldsymbol{X},\ \beta)$
for $s \in 1, \cdots, S$ **do**
draw $\psi_s \sim$ `LERW`$(\tilde{G}, \gamma_0, \gamma_k)$
weight by $w_s = \frac{1}{\det\ L_{(-i,-j)}(\tilde{G}/\psi_s)}$
end
estimate $\hat{\mathbb{E}}_\Psi[\Pr(\Gamma = \Psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)]\ =$
$\frac{1}{\sum_{s=1}^{S} w_s} \sum_{s=1}^{S} w_s \Pr(\Gamma = \psi_s \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)$
return $\hat{\mathbb{E}}_\Psi[\Pr(\Gamma = \Psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)]$

Procedure `LERW`$(\tilde{G}, \gamma_0, \gamma_k)$
initialize $\psi = (\gamma_0),\quad i = \gamma_0$
while $i \neq \gamma_k$ **do**
sample j uniformly from $\mathcal{N}_i$
step to $i = j$ and append to ψ
end
initialize $t = 0$
while $t < |\psi| - 1$ **do**
set t' to maximum index satisfying $\psi_t = \psi_{t'}$
if $t' > t$ **then**
erase elements in loop $(\psi_{t+1}, \cdots, \psi_{t'})$ from ψ
end
t+=1
end
return ψ

Algorithm 1: Approximating the probability that a random walk from γ_0 to γ_k is a path, up to the unknown multiplicative scaling factor $|\mathcal{P}|$, by importance-sampled Monte Carlo integration. The approximation converges to the exact likelihood as the number of simulations, S, approaches infinity. The loop-erasure proceeds along an unweighted random walk, identifies points where the walk returns to a previously visited node, then erases the second visit and all intervening nodes.

necessarily well-behaved, particularly when the network size is small. Thus, numerical optimization of the likelihood (with confidence intervals by the delta method with numerical Hessian) is inadvisable for short paths. When the parameter space is low-dimensional, the simulated likelihood can simply be evaluated on a fine grid. As the number of parameters increases, this approach rapidly becomes infeasible. In appendix A.3, I develop a procedure to estimate RPM by Metropolis–Hastings (MH) and discuss further approximations that can greatly speed computation. In appendix B, I use simulations to evaluate the proposed estimation procedure in various scenarios. Consistency is shown to depend not only on the number of paths, but also their length (and indirectly, network size).

4 U.S. Interstate Highways

I apply the RPM to the U.S. Interstate Highway System, often called the "greatest public works project in history." The interstate highways, estimated to cost over 500 billion inflation-adjusted dollars, provided the first comprehensive national road network for national defense and economic development. It offers an ideal empirical testing ground in that the planning process was highly transparent and explicitly stated decision criteria are publicly available. Interstate paths are analyzed with the RPM and alternative spatial models to assess whether these models accurately recover the decision-making process. The random-path model produces estimates that are consistent with planning criteria, whereas alternative spatial methods yield conflicting results.

4.1 Qualitative Benchmark

The Interstate Highway System evolved over several decades and numerous iterations. The 1921 Pershing map—an Army proposal of important routes for military logistics, including emergency mobilization—was one of the first comprehensive drafts. The counties to connect were identified in a report by the National Interregional Highway Committee (1944), along with a detailed

discussion of the route selection process. At this level, planners were fairly insulated from political pressures: An independent committee devised and applied a rule-based system, using census data, to identify counties to connect in each region. The committee was composed of professional bureaucrats—planners, civil engineers, and administrators from the Bureau of Public Roads. Legislative influence was primarily through a formula that fixed the distribution of funds among states, rather than influencing specific routes. Thus, the technocratic priorities described in this document represent a qualitative benchmark that can be used to test whether the RPM correctly models the decision-making process. The locations of highways within counties, in terms of the specific tracts of land to condemn, were proposed by the U.S. Bureau of Public Roads (1955) but are not considered here.

The highway committee states that "the recommended interregional system conforms closely" with military priorities, with extensive additions. The report discusses the specifics of route selection in detail, with the vast majority of attention devoted to connecting major urban centers. Planners started with a list of cities that had a 1940 census population over 100,000;[6]; between these, "the primary purpose was to select routes... which would join the principal centers of population and industry... by lines as direct as practicable." Thus, the Interstate Highway System was first and foremost designed to connect population centers, while keeping the overall system at a manageable length—in the ultimately recommended plan, just under 34,000 miles. While it was considered desirable to connect counties with high manufacturing capacity, planners observed that in practice, this goal could be achieved by maximizing the urban population served.[7]

Rural population was also described as an important consideration, with the proposed highways passing "en route between these hubs, through or very close to the denser clusters of population in

[6]Virtually cities with more than 100,000 residents were directly served, with three exceptions that were "passed in close proximity." The report explains that these cities could not be directly served without negatively impacting much larger adjacent cities.

[7]Roughly two-thirds of variation in 1939 manufacturing can be explained by population alone. From the committee report: "While slight differences exist in the relative importance of cities when they are measured on the one hand by their populations and on the other by the values added by their manufactures, on the whole the similarity of the measures is marked... It is, therefore, concluded that the recommended system closely approximates the system of optimum extent from the standpoint of service to manufacturing industry."

small towns and populous rural areas."[8] Agricultural production, on the other hand, was described as something of an afterthought, though the proposed system was shown to be adequate for the purposes of transporting farm products.

Just as importantly, planning documents describe a number of factors that were not important in route selection. Perhaps surprisingly, topography was an influence in "remarkably few places," and soil quality was not considered at all. Nor were interstates built along existing routes. Except for a few sections in the Northeast and Detroit, existing roads generally did not meet standards for lane width and arrangement, and "existing rights-of-way are grossly insufficient to permit such widening." These statements help justify the relatively simple model specifications used here.

4.2 Data

Following the National Interregional Highway Committee, I define the county as the unit of analysis. Based on National Highway Planning Network shapefiles, I convert a total of 57 two-digit interstate highways into sequences of adjacent counties.[9] I condition on the endpoints of these highways, as well as intermediate cities with populations over 100,000.[10] For example, I assume that I-5 was built as a route to connect Seattle to Los Angeles, with mandatory stops in Portland and Sacramento, but that planners were otherwise free to choose the intermediate steps. The interstates are thus split into 136 highway segments with fixed endpoints, passing through an average of 11 counties each. The highway system can be represented as a network in which nodes are highway counties and edges are dyadic highway connections. This highway network is a subgraph

[8]The report continues, "Indeed, the courses of the recommended routes are shown by this map to be in most instances the inevitable selections, if service of population is to be considered important in the choice."

[9] In the Interstate Highway System numbering scheme, long-range east-west (north-south) highways are given even (odd) two-digit codes, such as I-90 (I-95). Three-digit auxiliary highways that start with an odd number are spurs (e.g., I-391) and those that start with an even number are circumferential highways ("ring roads," e.g., I-495). Auxiliary highways are discarded.

[10]This is consistent with the planning process, in which routes were selected so that virtually all cities with populations over 100,000 were directly served. For computational reasons, two additional fixed waypoints were added—I assume that I-90 had to pass through Sioux Falls, SD and Buffalo, WY (junctions with I-29 and I-25, respectively). This was done to break up a highway segment of I-90 that was otherwise an extreme outlier in terms of length.

of the overall U.S. county network (shown in figure 1), because highway counties are a subset of all counties, and county-dyads with a direct highway connection are a subset of all geographically adjacent counties.[11]

Highway construction is modeled as a collection of random paths. Edge weights for the move from i to j are modeled by a softmax function with covariates based on planning documents:

$$\varepsilon_{i,j} = \frac{\exp\left(z_{i,j}\right)}{\sum_{j' \in \mathcal{N}_i} \exp\left(z_{i,j'}\right)}, \text{where}$$

$$z_{i,j} = \beta_{\text{dist}}\texttt{dist}_{i,j} + \beta_{\text{pop}}\texttt{pop}_j + \beta_{\text{urb}}\texttt{urb}_j + \beta_{\text{mil}}\texttt{mil}_j + \beta_{\text{ind}}\texttt{ind}_j + \beta_{\text{agp}}\texttt{agp}_j$$

and covariates are defined as follows (census data from Haines, n.d.):

- $\texttt{dist}_{i,j}$: Minimum road distance between county seats of i and j, miles.[12]
- $\texttt{pop}_j$: 1940 log census population of county j.
- $\texttt{urb}_{j,m}$: 1940 urban census population of county j. Discretized into 5 dummy variables with breakpoints at 2, 10, 25, and 50 thousand, following planning documents.
- $\texttt{mil}_j$: Military facilities in j, as proxied by log spending from 1940–1945 reported by the Civilian Production Administration (CPA).
- $\texttt{ind}_j$: Industrial capacity of county j, as proxied by CPA-reported log total value of industrial facility expansions from 1940–1945.[13]
- $\texttt{agp}_j$: 1939 log value of agricultural products sold and traded in county j.

[11] Two counties, A and B, are considered adjacent if (1) they share a border or (2) their county seats can be connected by a line that barely clips a third county, C. In a hypothetical set of 10 mi. $\times$ 10 mi. counties forming a square grid, the latter criteria allows for diagonal connections between counties. The threshold for "barely clips" is arbitrarily defined as one-quarter of county C's characteristic length (the square root of its area). For example, in the square grid described above, a road can clip at most a triangular region of area 2.5 sq. mi. (out of C's total area of 100 sq. mi.) before it is considered an $A - C - B$ path, rather than an $A - B$ path.

[12] For approximately "adjacent" counties that do not share a border, I first identify the points at which i's border is closest to j's border. The minimum road distance is then defined as the distance from i's seat to the point on its border, plus the minimum distance between i and j's border, plus the distance from the point on j's border to its county seat.

[13] Results are substantively and statistically indistinguishable when using 1939 log manufacturing value added, but this earlier data has substantial missingness in non-urban aras

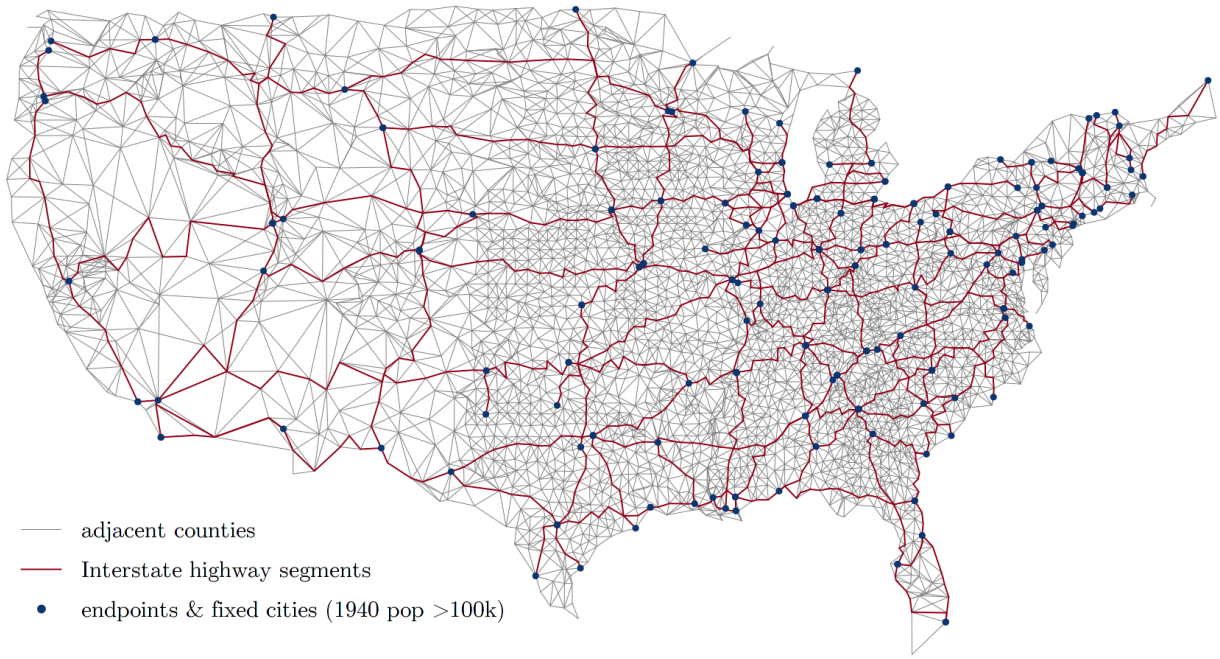

Figure 1: The graph of all adjacent counties is drawn in thin gray lines. The subgraph of adjacent counties connected by Interstate Highways is highlighted with thicker red lines. The subgraph is composed of paths between waypoints, plotted as blue circles.

4.3 Results

Based on observed decisions by planners as they connect major cities through a series of intermediate counties, the random-path model estimates the priorities of the Interstate Highway System. For example, figure 2 shows a case in which Interstate 80 deviates from the shortest possible route between Cheyenne, WY and Omaha, NE. While such cases appear to suggest that planners are willing to trade off some additional distance in order to pass through small cities and military facilities, visual inspection alone cannot determine whether the relationship is statistically significant, nor assess the value placed on cities relative to military bases.

Because interstates are bidirectional but the RPM analyzes directed paths, I treat each segment as the equally weighted combination of both directions, e.g., the northbound and southbound parts. Parameter posterior distributions and convergence diagnostics are reported in appendix C

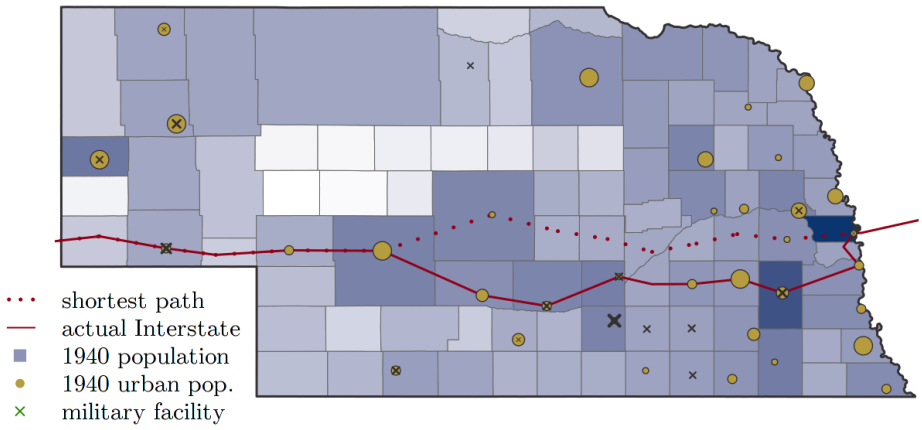

Figure 2: The I-80 segment between Cheyenne, WY and Omaha, NE (solid red line) deviates from the shortest route (dotted red line). Rather than minimizing distance, highway planners opted to connect small cities (yellow circles) and military facilities (black crosses).

For interpretability, parameter estimates are converted to a probability scale in figure 3 according to the following scenario: Suppose a highway comes upon a mutually exclusive choice between two identical counties, j and j'. Holding the rest of the route fixed, the highway has an equal probability of passing through either. If j was changed so that it had some desirable property, such as a higher population, what would be the corresponding increase in the probability of highway construction through it?

RPM results suggest that planners preferred more direct routes. They reveal a preference for connecting counties with higher population, particularly when this population was concentrated in cities. The addition of an average manufacturing facility or a one-standard-deviation increase in agricultural production was found to be small and insignificant, whereas the addition of an average military facility was associated with a large increase in the chances of receiving an interstate. These model predictions agree well with the qualitative priorities outlined above (according to the National Interregional Highway Committee, routes were to be "as direct as practicable" and "close to the denser clusters of population," with "close proximity of... [military and naval] establishments to the recommended routes").

In contrast, alternative spatial models with identical specifications (except distance between

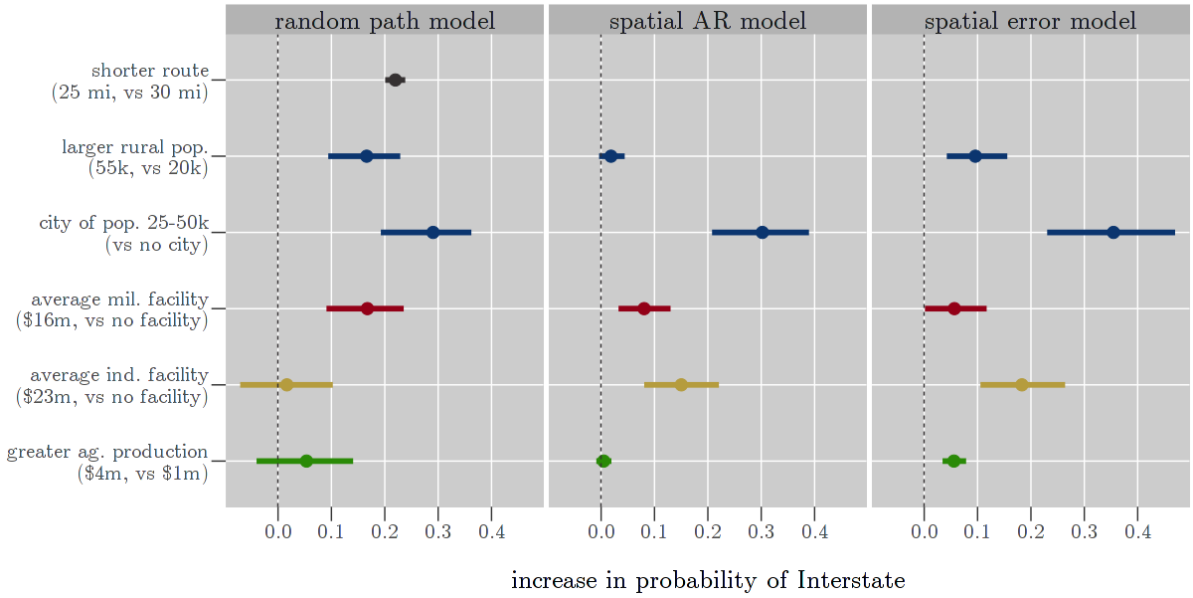

Figure 3: Predicted increase in probability of highway construction in one of two otherwise identical counties, holding all else fixed. Base distance of 25 miles is roughly the median distance between county seats connected by interstate highways. Various changes in county attributes (y-axis) are approximately one standard deviation in the attribute, except for military and industrial capabilities, which represent the value of an average facility. Points are posterior means; error bars are 95% posterior credible intervals.

counties, a dyadic covariate that cannot be incorporated) produce significantly different estimates that do not agree with documented interstate priorities. A standard probit model with state fixed-effects and a correction for spatially correlated errors incorrectly suggests that an average industrial facility is more important than an average military facility, by a large and statistically significant margin—a surprising and almost certainly incorrect result, given the planning process (and the fact that the system is named the "National System of Interstate and Defense Highways"). The spatial autoregressive probit model finds that if city size is held fixed, a one-standard-deviation increase in total population has no substantive effect on highway construction. This directly contradicts planning documents (from the National Interregional Highway Committee, 1944, "the recommended routes trace their courses along the country's most populous bands [of rural population]... the evidence of appropriate selection is marked").

The failure of alternative spatial models is due to two reasons. First, the path structure

of highways is a violation of the underlying assumptions of these spatial models. Second, and more importantly, by nature they cannot account for the distance between sequential highway counties—perhaps the single most important factor in route choice. These alternative models do not capture how highways deviate from the shortest route to touch desirable counties. Instead, they essentially compare highway counties to non-highway counties, ignoring the fact that many rural areas were never viable candidates for an interstate.

5 Navigating the Streets of Baghdad

In Baghdad, the relationship between majority Shia and minority Sunni Muslims has been one of peaceful coexistence for centuries (Tripp, 2000). This history of integration and intermarriage stands in stark contrast to a wave of ethnic cleansing in 2006–2007 which has dramatically reshaped the city's ethnic landscape (Baker et al., 2006). What are the long-run effects of this civil conflict—and the resulting counterinsurgency campaign—on Baghdad's political geography? The prospects for post-conflict reintegration hinge on whether these changes have been driven solely by *need-based* sorting (Morrison, 1993) or whether they have also been accompanied by the emergence of *taste-based* sorting (Schelling, 1969, 1971).

There is no dispute that need-based sorting—moving out of mixed neighborhoods to escape violence—was an important factor in Baghdad's sudden segregation. During the 2006–2007 conflict, sectarian militias drove many families out of their houses at gunpoint, and others were intimidated into moving preemptively. What is less clear is whether conflict fundamentally changed ethnic relations in Baghdad and led to taste-based sorting. If the conflict was simply a power struggle between armed factions, such as Ba'ath loyalists and the Mahdi Army, citizens' preferences might remain unchanged. On the other hand, conflict could have precipitated a shift in citizens' ethnic attitudes. In this case, even people who were unaffected by violence would move out of mixed-sect areas due to a newfound distaste for old neighbors.

This question is not merely of historical interest. Out-group aversion is an important parameter in game-theoretic models of segregation, and different values lead to very different predictions about the future trajectory of Baghdad or other cities segregated by violence (Young, 1998; Zhang, 2004*a*,*b*). If sorting was a need-based response to conflict, then segregation is not stable: After the conflict ends, the city will eventually return to its former integrated state. If there is a substantial taste-based component, however, the current geographic division of Sunnis and Shia is likely to persist far beyond the end of conflict.

Using the RPM, I analyze behavioral data from a field activity by Christia and Knox, in which subjects participate in a "treasure hunt" in their own home neighborhoods. Results suggest that ethnic conflict over the past decade has led to the emergence of previously nonexistent taste-based sorting. In one scenario—a hypothetical one-kilometer walking task that can be completed in about 12 minutes—estimates indicate that Shia will go out of their way by 34% (4 minutes) to avoid Sunni areas, and Sunnis will go out of their way by 12% (1.5 minutes) to avoid Shia areas. These shifting societal preferences have implications not only for reintegration, but also for economic and political development as the Iraqi state attempts to rebuild after years of conflict.

5.1 Theory and Background

The geographic impact of conflict on Baghdad residents, shown in figure 4, is hard to overstate. Sunni and Shia once lived side-by-side in nearly every district of the city, but formerly mixed areas are now overwhelmingly dominated by one sect or the other. What drove the sudden segregation in 2006–2007? Clearly, need-based sorting in response to sectarian purging was an important factor. The effects of violence on migration are examined by Morrison (1993), who developed a model that incorporates preferences for safety in addition to economic considerations such as wage maximization. But did the aftershocks of this conflict also change sectarian attitudes and lead to taste-based sorting, where citizens moved due to a newfound aversion to their out-group neighbors? The answer to this question matters, because violence fades—recent deaths, though

still substantial, have fallen below half the peak in late 2006 (Iraq Body Count, 2016). Ethnic attitudes, on the other hand, can persist for years if not decades.

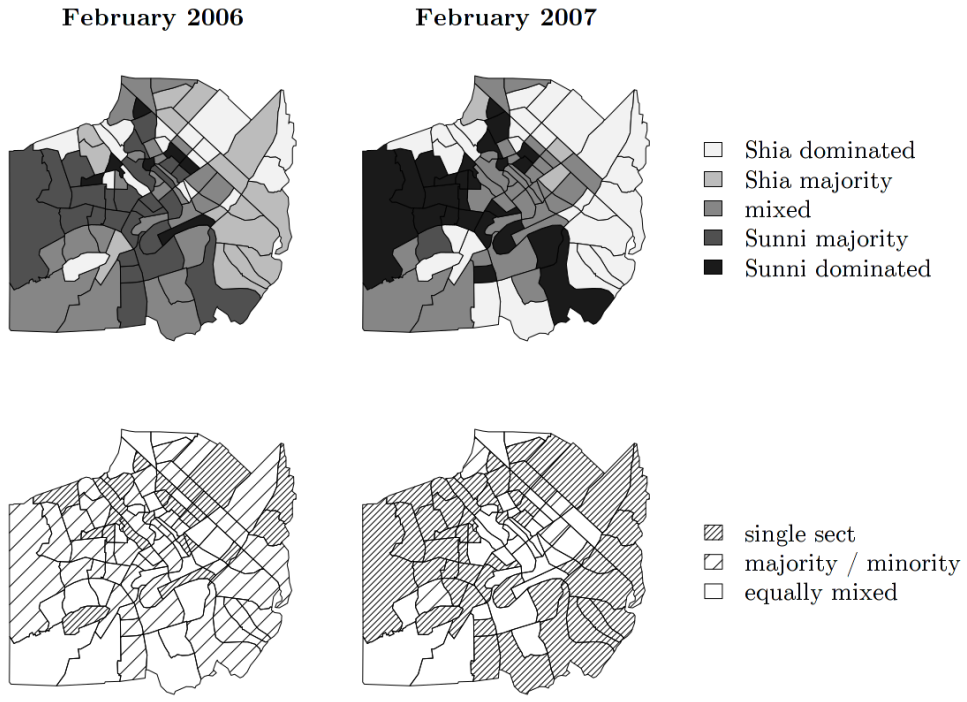

Figure 4: Top panels show sectarian composition of Baghdad pre- and post-purges (adapted from International Medical Corps, 2007). Darker (lighter) districts have a higher proportion of Sunni (Shia). Bottom panels show the sectarian diversity of districts; denser hatch marks indicate less diverse areas. As a result of local sectarian cleansing, areas that were previously mixed with a Sunni majority (dark gray, slight hatching) tend to become all-Sunni (black, dense hatching), and those that were previously mixed with a Shia majority (light gray, slight hatching) become all-Shia (white, dense hatching).

The daily lives of Baghdadis can help shed light on these questions. In terms of daily movement, need-based sorting suggests that as citizens navigate their surroundings, their decisions are driven primarily by a desire to survive, earn a living, and procure food. Citizens will generally act efficiently, taking direct routes when walking, except when facing the threat of violence. If Baghdad's segregation is predominantly a need-based response to this violence, then we may expect gradual desegregation after conflict ends: Citizens will stop avoiding out-group areas as they become safer, and families will move back into former homes or sort into new neighborhoods (Tiebout, 1956), with the city eventually shuffling back to an integrated state. Although the claim

may seem implausible in the current climate, integration has proven to be surprisingly robust in the past. Baghdad did not segregate in response to Shia and Kurdish uprisings in 1991 or the brutal suppression that followed, even after a major uptick in sectarian tensions. To the extent that increased contact can slowly reduce ethnic animosity in post-conflict settings (Samii, 2013; Mironova and Whitt, 2014; Hartman and Morse, n.d.), this scenario suggests the possibility of reconciliation.

Taste-based sorting, on the other hand, implies that citizens will pay costs to avoid the discomfort of out-group contact even when there is no threat to safety or economic rationale. For example, it predicts that they will walk far out of their way to avoid out-group areas. If true, this would be a new development in Baghdad's history of cosmopolitanism. Even after the fall of Saddam Hussein's dictatorship in 2003, when restrictions on movement were abolished, there was little to no change in Baghdad's ethnic composition—Sunni and Shia continued to live in close proximity. After the 2006–2007 purges, however, there are some reasons to suspect that this may be changing. For one, institutional changes have increasingly elevated the prominence of sectarian identity. Moreover, the counterinsurgency strategy of coalition forces has been one of divided political geography.[14] Coalition-built walls and armed checkpoints intended to protect single-sect neighborhoods have been criticized for "hardening the separation of Sunnis and Shias," with effects that could persist long after their ongoing dismantlement (Damluji, 2010). Research starting from Schelling (1969, 1971) shows that these conditions can lead to a rapid "tipping point," beyond which segregation is hard to escape. If political violence has contributed to or been accompanied by the rise of taste-based sorting, these agent-based and game-theoretic results suggest that society is locked in a long-term segregated state (Young, 1998; Zhang, 2004*a*,*b*).

Empirically testing these competing hypotheses is a difficult task. Surveys offer one proxy of sectarianism but are often subject to social desirability bias. For example, in a separate survey of religious Shia pilgrims, Christia, Dekeyser and Knox (n.d.) show that pilgrims claim to support

[14]As an example, the well-trafficked Bridge of the Imams, which served as a point of contact between Sunni Adhamiya and Shia Kadhimiya, was barricaded for years to stop armed conflict between the neighborhoods.

Sunni–Shia interaction but still overwhelmingly favor co-ethnic neighbors in a conjoint experiment (Hainmueller, Hopkins and Yamamoto, 2014).[15] While indirect survey methods can accurately measure sensitive attitudes and even certain kinds of past behavior, such as sensitive vote choices (Rosenfeld, Imai and Shapiro, 2016), they are unlikely to perform well in this particular context. This is because predictions about re-integration hinge on future behavior. Indirect methods that tap into sectarian attitudes do not directly address this question, since not all biased individuals will express their attitudes through costly and publicly visible behavior; moreover, survey questions that directly ask about future behavior in hypothetical scenarios are notoriously unreliable even for non-sensitive questions (Rogers and Aida, 2013).

Instead, I draw on behavioral data from Christia and Knox (n.d.) that isolates the taste-based channel through a field activity. We assign matched Sunni and Shia residents to a treasure hunt through mixed-sect areas in their home districts. By assigning participants to find the same target locations, we sidestep a fundamental problem in observational behavioral data: If people live in neighborhoods where all basic needs are met, it might appear that they avoid surrounding areas—including those populated by out-groups—but this would not indicate an unwillingness to move to those areas after violence dies down. In our field study, locations are carefully selected, with input from a mixed-sect team of local advisors and officials intimately familiar with the area, to eliminate any reasonable concerns that participants might have about their personal safety. This avoids a second confounder of taste-based preferences in observational data—that the threat of violence (generally unobservable, unless local knowledge is available) may be associated with out-group areas.

[15] Among these religious respondents, the only trait less desirable than Sunni faith was alcoholism, which is seen as a serious moral failing in Iraq.

5.2 Sample and Design

To test these models of sorting, we recruited a group of University of Baghdad students from mixed-sect districts to participate in a field navigation activity—a treasure hunt. These recruits are not representative of Baghdad as a whole. Instead, they represent a subpopulation in which taste-based avoidance is "least likely" to be found (Eckstein, 1975; Gerring, 2007). If taste-based aversion exists even among well-educated students who attend a mixed-sect university and live in low-conflict, diverse areas, we may safely conclude that it is a widespread phenomenon.

We advertised around campus for students living in two districts, Ghazaliya and Jihad, chosen for their sectarian diversity and security (shown in figure 5). While both districts are mixed-sect, their sectarian landscapes differ in important ways—Ghazaliya tilts toward Sunnis and Jihad has a larger Shia population, though both districts contain substantial numbers of each sects. Central Ghazaliya has neighborhoods in which Sunni and Shia live side-by-side, but Sunnis (Shia) tend to spend their time on commercial streets in the Sunni-dominated (Shia-dominated) areas to the south (north). In contrast, neighborhoods in Jihad tend to be single-sect, but both groups frequent markets and cafes on the same major thoroughfares.

All potential recruits provided basic demographic information in an initial meeting, and 120 Sunni and Shia participants were chosen so that groups were comparable in terms of on gender, district, and household income sufficiency. The average participant was 21 years old. Our gender balance was skewed towards male participants (two-thirds), and 55% of participants reported that household income was sufficient to cover costs. By design, there were no significant differences between sects (see table 1).

The field navigation activity was embedded in a broader week-long smartphone study in which participants consented to our collection of behavioral data on social networks, traditional and social media consumption, and location. To incentivize participation, subjects were given a recent-model Android smartphone to use with their own SIM card for the duration of the study. They received a free one-month credit (covering the study period and three additional weeks) for free

data, domestic calling and text messaging. In addition, subjects could earn up to 15,000 IQD in phone credit for completing the treasure hunt (about 13 USD, c.f. laborer day wages of 7–30 USD depending on skill, or civil service monthly wages of 500 USD). This was a substantial amount for participants and was seen as highly motivating. Smaller amounts of phone credit were offered for other, shorter experimental tasks. After the end of the study, over half of the participants won a contest that allowed them to keep their phone.

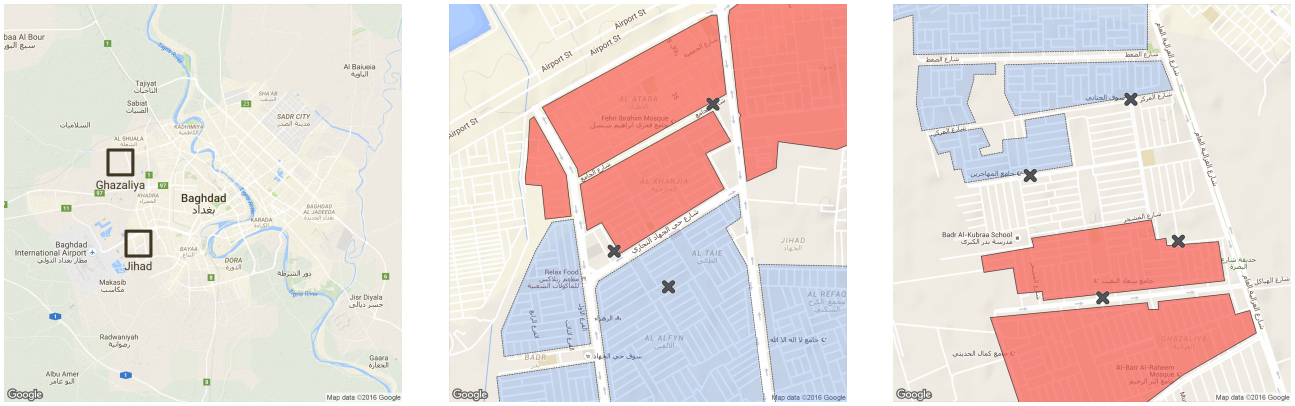

Figure 5: Left panel indicates location of Jihad and Ghazaliya in Baghdad. Center panel shows the treasure hunt area in Jihad, with targets (black ×) and Sunni (dark red) and Shia (light blue) residental areas. Neutral commercial streets and or mixed residential areas are not marked. Most Jihad participants walked between the western, northern, and southern targets in that order. The right panel shows the Ghazaliya playing field. Most Shia participants walked from the northern target to the western one, then finished at one of two southern targets; most Sunnis did the route in reverse order.

In each district, we chose centrally located targets that were near to both Sunni- and Shia-dominated areas. These were well-trafficked locations, such as markets, bus stations, schools, and mosques. Locations were selected from a set of options that local advisors of both sects agreed would eliminate reasonable concerns about security. The vicinity of each treasure hunt was an area visited by people of both sects at least occasionally. Besides avoiding harm to participants, this also helped ensure that differences in walking routes were due to taste-based aversion, rather than need-based concerns about safety.

Each participant was assigned to a supervisor who provided treasure-hunt directions via an

instant messaging app. Supervisors were instructed to initially provide a start location, but no further information. After a participant arrived and verified their location with a "selfie", they would receive the next location. A total of two additional locations were assigned, so that the shortest possible route would be about three kilometers long (30 minutes) and pass through residential neighborhoods of both sects. The resulting route consists of two paths: from start to midpoint, and from midpoint to endpoint. These locations are shown in figure 5, along with nearby sectarian neighborhoods.

In a follow-up survey, walkers reported their familiarity with the area and whether they stopped to ask for directions in out-group areas, among other questions. To supplement our phone-based location data collection, walkers also self-reported their routes by drawing on a map of the treasure hunt area.

5.3 Nonresponse, Compliance, and Attrition

We gathered a total of 102 paths from 55 unique participants, containing an average of 21.1 correlated decisions per path.[16] Nonresponse was high, in large part due to extremely low participation by women (30 percentage points lower than men, $p = 0.01$). This pattern was predictable, given local gender norms. We were aware that treasure-hunt nonresponse would be higher among women, but chose to keep both genders in the study so that they would not be excluded from social media and other experimental modules. Other observable characteristics—sect, home district, age, or income sufficiency—are uncorrelated with nonresponse (participation and completion rates are given in table 2). Based on conversations with supervisors, subjects' participation in the treasure hunt appears to be largely driven by whether someone was willing to exercise for 30–60 minutes. However, we cannot rule out the possibility that subjects who are more averse to out-groups are

[16]Five of these paths are instances where the participant returned to the starting point (thus adding a third leg) or revisited the treasure-hunt area at some point over the weekend and happened to walk a different route between the same locations. The latter is unsurprising since our locations are commonly visited by locals. I include these in the analysis because they provide additional information about walking patterns.

also less likely to participate.[17]

Compliance with the task protocol was imperfect. A handful of early participants were accidentally informed of all targets at once, allowing them to walk the route in a different order than intended. In Ghazaliya, other participants visited a different grocery store than originally intended, due to ambiguous instructions. I address these deviations by assuming that they are independent of out-group aversion, then conditioning on the start- and endpoints of each segment. If participants deliberately choose one store to avoid certain areas, the violation of this assumption will bias estimates toward zero.[18] Participants were free to withdraw at any time, and 13 failed to complete the treasure hunt despite succeeding in the first leg. This attrition was distributed evenly by participant's sect or home area. Anecdotally, it seemed to occur when participants learned that their next target was in an out-group area. This suggests that more-averse participants are more likely to withdraw, which will bias estimates toward zero. There is some evidence that lower-income participants have a lower attrition rate ($p = 0.085$), which could be due to the financial incentive for completion. In general, however, reactions to the task were quite positive. Typical responses in the debrief survey conveyed excitement ("an enjoyable new experience", "I felt adventurous", "I liked exploring") or discussions of the physical exertion ("tiring but entertaining", "good for health", "healthy exercise").

[17] One way this might happen is if highly averse people are also less mobile or active, perhaps due to fear. It might also arise if participants were aware of the sectarian nature of the task before starting, although steps were taken to prevent this. Walkers were asked not to discuss the task with other subjects, and they were unaware that all participants were assigned an identical set of targets in a neighborhood. In addition, in open-ended responses from debrief surveys, we saw no indication that subjects realized the treasure hunt was intended to send them to out-group areas. However, subjects were allowed to complete the treasure hunt at any time over a two-day period, and we cannot rule out the possibility that later participants might be aware that some targets are located in out-group areas.

[18] If noncompliance is associated with higher levels of out-group aversion—for example, if more-averse subjects walk in the wrong order because it allows them to avoid out-group neighborhoods more easily—then conditioning on this decision will result in attenuation bias in estimates of out-group aversion.

	Sunni mean	Shia mean	p-value
Ghazaliya	0.72	0.62	0.36
Jihad	0.28	0.38	0.36
Age	20.9	21.3	0.21
Male	0.71	0.60	0.30
Income sufficient	0.57	0.53	0.87
N	62	58	

Table 1: Summary statistics for Sunni and Shia subjects, with p-values from t-test and chi-squared tests for continuous and binary variables, respectively.

	Participated in activity	p-value of difference	Completed both legs	p-value of difference
Sunni	0.48	0.78	0.80	0.43
Shia	0.43		0.72	
Female	0.27	0.00	0.91	0.36
Male	0.56		0.73	
Ghazaliya	0.51	0.15	0.76	0.36
Jihad	0.36		0.79	
Income sufficient	0.52	0.38	0.83	0.08
Income insufficient	0.41		0.65	

Table 2: First and second columns describe response rates by subjects in various subgroups, with p-values for the difference based on a multivariate probit regression. Third and fourth columns describe completion (non-attrition) rates among subjects who started the treasure hunt.

5.4 Data

Christia and Knox wrote a custom Android app to record the location of participants at one-minute intervals, as well as the accuracy of the location estimate. Because this app was active for the entire week-long study, we used a mix of GPS- and Wi-Fi-based measurement to compromise between accuracy and power consumption. GPS estimates are typically higher quality, with an accuracy within 10 meters in outdoor areas, but we find that Wi-Fi based location is generally sufficient for our purposes (often on the order of 10 to 100 meters) and that the far lower power requirements of Wi-Fi counterbalances its lower accuracy for long-term tracking. Walking routes are constructed from location data as shown in figure 6. When smartphone location data is unavailable due to technical issues, I use self-reported routes that participants drew on a map

during debrief.[19]

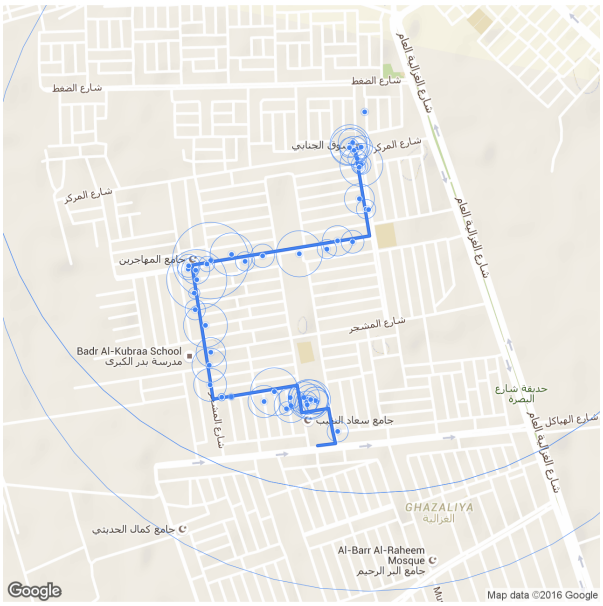

Figure 6: Example of smartphone measurement of walking route. Blue points represent estimated locations. A thin circle is drawn around each point, with a radius corresponding to the estimated accuracy. A path that aligns these points with the street network is manually fit to the location data.

I model treasure-hunt routes as random paths on a street network (modified from OpenStreetMap, 2016), where nodes are intersections and edges are street segments. Covariates are indexed as follows: i denotes the current node (the start of a step); j is the node to which the walker is moving (end of step); l represents a leg of the treasure hunt (e.g., from the mosque to the school); and k denotes an individual. Edgewise covariates that describe each street segment are coded based on satellite imagery and input from our mixed-sect team of local advisors who were familiar with the area.[20] Additional variables are taken from the debrief survey. These covariates are given below.

- $\texttt{dist}_{i,j}$: Length of street segment (1 = 100 meters).

[19] When both sources are available, they are generally consistent.

[20] Coders were given printed maps of the area, which they annotated by drawing borders around single-sect residential areas. Coders provided additional information about these areas, such as their safety level and professional composition. Major thoroughfares were traced and described, and major landmarks such as schools were marked. Annotated maps were then manually aligned and digitized.

- $\texttt{direct}_{i,j,l}$: Directness of approach, or how much closer the $i \to j$ step brings a walker to the endpoint of leg l (in units of 100 meters).
- $\texttt{enclosed}_{i,j}$: Indicator for narrow enclosed streets in densely built residential areas. Coded from satellite imagery.
- $\texttt{thoroughfare}_{i,j}$: If street is a major thoroughfare, e.g. a major commercial avenue. Coded from local knowledge and satellite imagery.
- $\texttt{sunni\_walker}_k$ and $\texttt{shia\_walker}_k$: Participant sect (binary).
- $\texttt{familiar}_{k,l}$: In debrief survey, whether participant indicated that they were familiar with the endpoint of leg l (binary).
- $\texttt{outgroup}_{i,j,k}$: Whether street passes through an area dominated by out-group residents (binary). Boundaries are based on local team's knowledge of neighborhoods.
- $\texttt{safety}_k$: In debrief survey, whether participant indicated that safety was a factor in their route choice (binary).
- $\texttt{familiar}_{k,l}$: In debrief survey, whether participant indicated that they were familiar with the endpoint of leg l (binary).

Examples of thoroughfares and enclosed neighborhoods are given in figure 7. Street networks are shown in figure 8 with these geographic covariates.

Figure 7: Annotated satellite imagery in Jihad, near starting point. Intersections are marked with dots, with lines depicting the connecting streets. Major thoroughfares (thick lines) cross near a bus station, at lower left. Streets that are fully enclosed by residential areas are drawn with dotted lines.

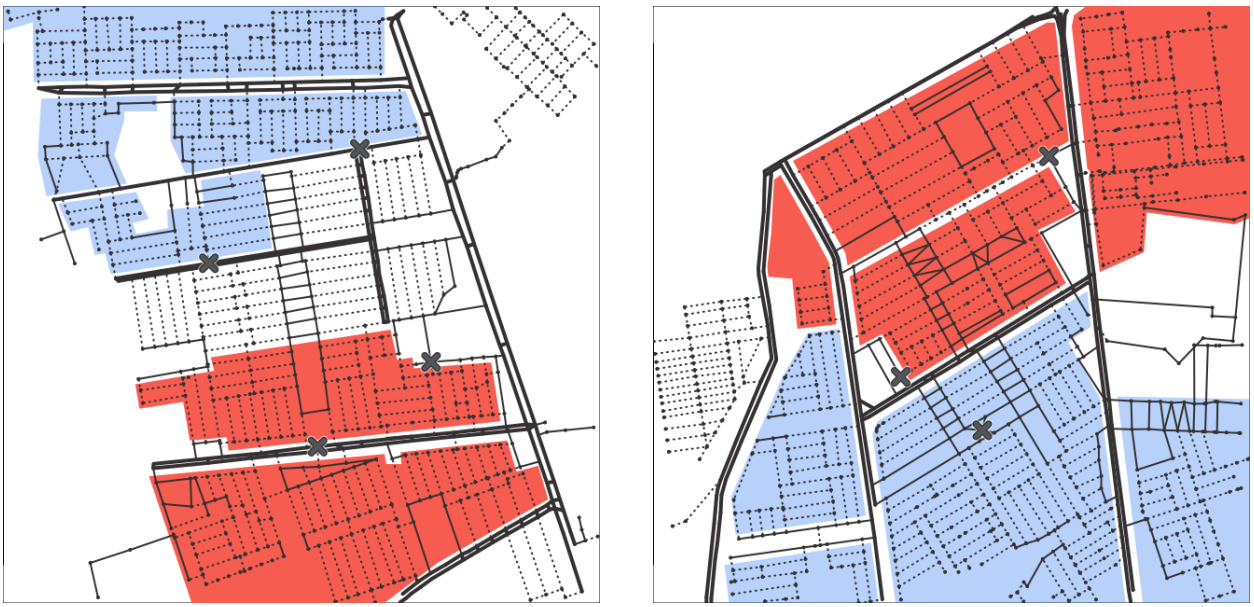

Figure 8: Street network in Jihad (left) and Ghazaliya (right). Targets are marked with a black $\times$. Sunni- and Shia-dominated residential areas are highlighted in dark red and light blue, respectively (neutral or mixed regions are left white). Major thoroughfares are indicated with solid thick lines, "open" streets in residential areas (e.g., bordering a park) are in solid thin lines, and "enclosed" streets in densely built residential areas are drawn in dotted thin lines.

5.5 Model and Results

I first describe the preferred model specification and describe the results from this model. I then test alternative explanations and robustness with other specifications. Estimates and 95% credible intervals for all models are given in figure 9. Finally, to interpret these results, I describe how the sectarian landscape affects Baghdadis' route choices in several simple but realistic scenarios.

5.5.1 Baseline Specification and Results

In the baseline specification, edge weights are modeled with the softmax function,

$$\varepsilon_{i,j,k,l} = \frac{\exp\left(z_{i,j,k,l}\right)}{\sum_{j' \in \mathcal{N}_i} \exp\left(z_{i,j',k,l}\right)}, \text{where}$$

$$z_{i,j,k,l} = \beta_{\text{dist}}\texttt{dist}_{i,j} + \beta_{\text{direct}}\texttt{direct}_{i,j,l} + \beta_{\text{enclosed}}\texttt{enclosed}_{i,j} + \beta_{\text{thoroughfare}}\texttt{thoroughfare}_{i,j} +$$

$$\beta_{\text{SunniOG}}\texttt{sunni\_walker}_k \cdot \texttt{outgroup}_{i,j,k} + \beta_{\text{ShiaOG}}\texttt{shia\_walker}_k \cdot \texttt{outgroup}_{i,j,k}.$$

Note that the base terms for `sunni_walker` and `shia_walker` are omitted, because their coefficients are statistically unidentified. (From the perspective of a walker standing at any node i, all step options would have an identical constant added to the linear predictor, C; this can be rewritten as a multiplicative constant e^C on both the numerator and denominator, then canceled.) Three MCMC chains were run with 10,000 iterations per chain; convergence diagnostics are shown in appendix D.

Results from the baseline model show that minority Sunnis prefer to avoid entering out-group (Shia) areas. For Sunni walkers, the coefficient on "enter out-group area" is negative and statistically significant, indicating that these streets are less likely to be selected. However, Sunni aversion is relatively small compared to Shia aversion: Members of the Shia majority are significantly more reluctant to enter out-group (Sunni) areas. These results show that even after eliminating need-based reasons to avoid out-group areas, Baghdadis still exhibit sectarian taste-based aversion. Moreover, taste-based aversion is significant even among what is perhaps the best-integrated subpopulation in Baghdad, suggesting that it is likely stronger among the rest of society.

Other estimates are generally intuitive. Participants prefer shorter routes (negative coefficient on distance), avoid dense residential developments, and prefer to walk on major thoroughfares. A null result was found for the directness of route. This may be because directness and distance essentially measure the same concept—a walker who takes the shortest route to a destination is also moving in the correct direction—but in principle, including both terms leads to a more flexible specification because walkers may respond to long steps differently under certain circumstances, e.g. when they overshoot the destination. Findings for these covariates correspond well with anecdotal evidence from participants, who reported that thoroughfares such as commercial streets were "exciting," with more activity and shops, and that they "had no reason" to cross through

residential areas where they "did not belong." Coefficient estimates for the basic specification are interpreted in section 5.5.2, and alternative specifications are discussed in section 5.5.3.

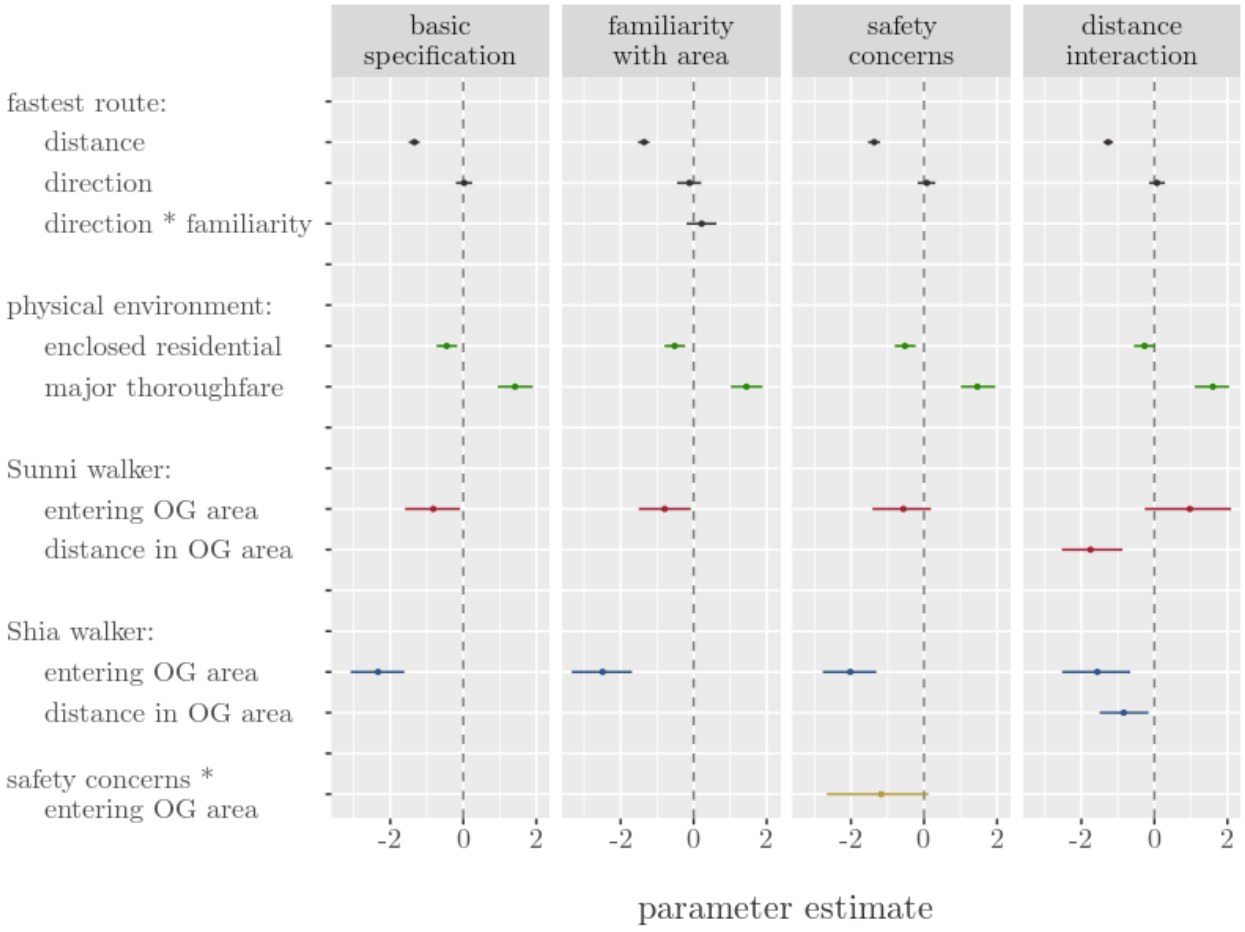

Figure 9: Results from all model specifications. Points are posterior means; error bars are 95% posterior credible intervals. Distance is measured in units of 100 meters.

5.5.2 Interpreting Results

Coefficient estimates in figure 9 can be interpreted relative to each other. For example, the negative coefficient on `distance` is roughly the same size as the positive coefficient on `thoroughfare`, so the "cost" of walking 100 additional meters could be offset by the "benefit" of staying on a major thoroughfare. However, this sort of interpretation paints an incomplete picture of the iterative decision-making process that RPM is designed to model.

Instead, consider a hypothetical task in which a walker must decide whether or not to cross through an out-group area to get to a destination that is one kilometer away. Figure 10 presents two versions of this task. In the first scenario, both routes are equal in length. The corresponding estimates in the lower panel (leftmost error bars) show that both Sunni and Shia are significantly more likely to avoid the out-group route when it is costless (no additional distance required). Sunni will cross through Shia territory 31% of the time (significantly less than the 50% chance if sect were irrelevant), and Shia will choose to cross through Sunni areas only 9% of the time. In the second scenario, the out-group route is shorter, and the alternative is twice as long. In this case, the rightmost error bars show that both Sunni and Shia would almost certainly cut through out-group areas to save a kilometer of additional walking, as in the second scenario. Estimates in the middle show how these decisions change as the distance tradeoff shifts between these two extremes: Sunni become exactly indifferent between the route options when out-group avoidance "costs" 120 meters, or roughly a one-eighth increase in walking time (95% credible interval [10m, 240m]).[21] Shia are more reluctant, becoming indifferent at 340 meters—about a one-third increase in walking time ([240m, 460m]).

Figure 11 demonstrates the estimated behavior of participants in more complex scenarios, when many options are available, by simulating 1000 walking routes using point estimates of the RPM parameters. These plots show that in aggregate, members of the Sunni minority will tend to deflect slightly away from the most direct route to avoid Shia areas, but many individuals are willing to pass through. In contrast, Shia walkers stay just outside the border of the Sunni area, with almost no individuals cutting across. Members of both sects will go far out of their way to use a major thoroughfare, instead of walking through residential areas.

These results illustrate a strength of the RPM—that they illuminate how changes in short-term incentives can have broader implications for long-term behavior. Unlike existing alternatives, it is explicitly designed to model the path formation process. Thus, researchers can use the model as a

[21] For added distances in this range, the route choice probability is not significantly distinguishable from 0.5 at the 95% credible level.

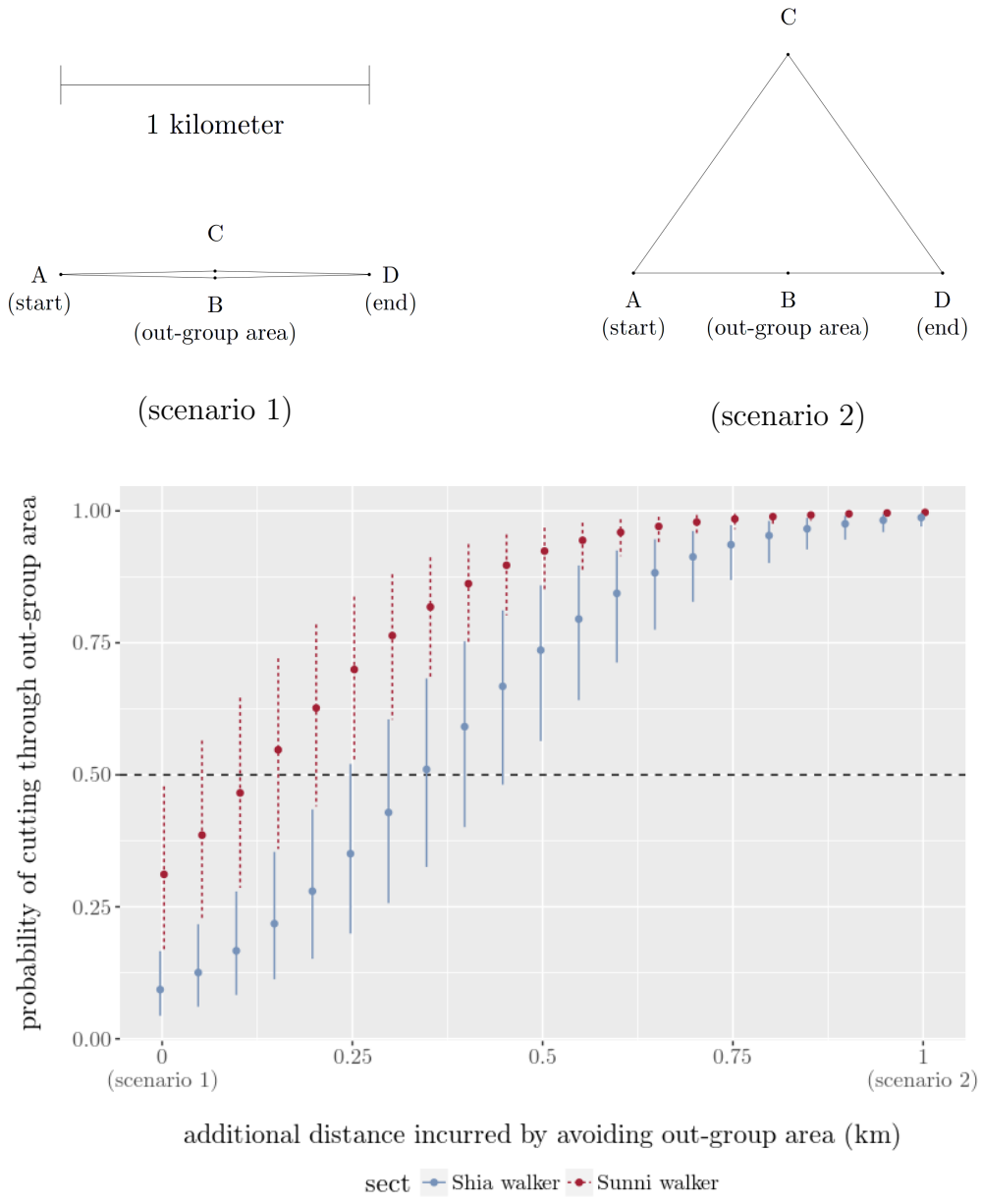

Figure 10: Top panels depict scenarios in which a walker must choose between two potential routes. $A - B - D$ passes through an out-group area, while $A - C - D$ does not. In scenario 1, these routes are of equal length, so there is no incentive to pass through the out-group area. The corresponding estimates (leftmost error bars) show that under these conditions, both Sunni and Shia significantly avoid the out-group route. In scenario 2, $A - C - D$ is twice as long as $A - B - D$, so walking through the out-group area saves one kilometer. The rightmost error bars show that this provides a sufficient incentive for both Sunni and Shia to overcome their aversion. Estimates in the middle show results for various intermediate scenarios between these extremes.

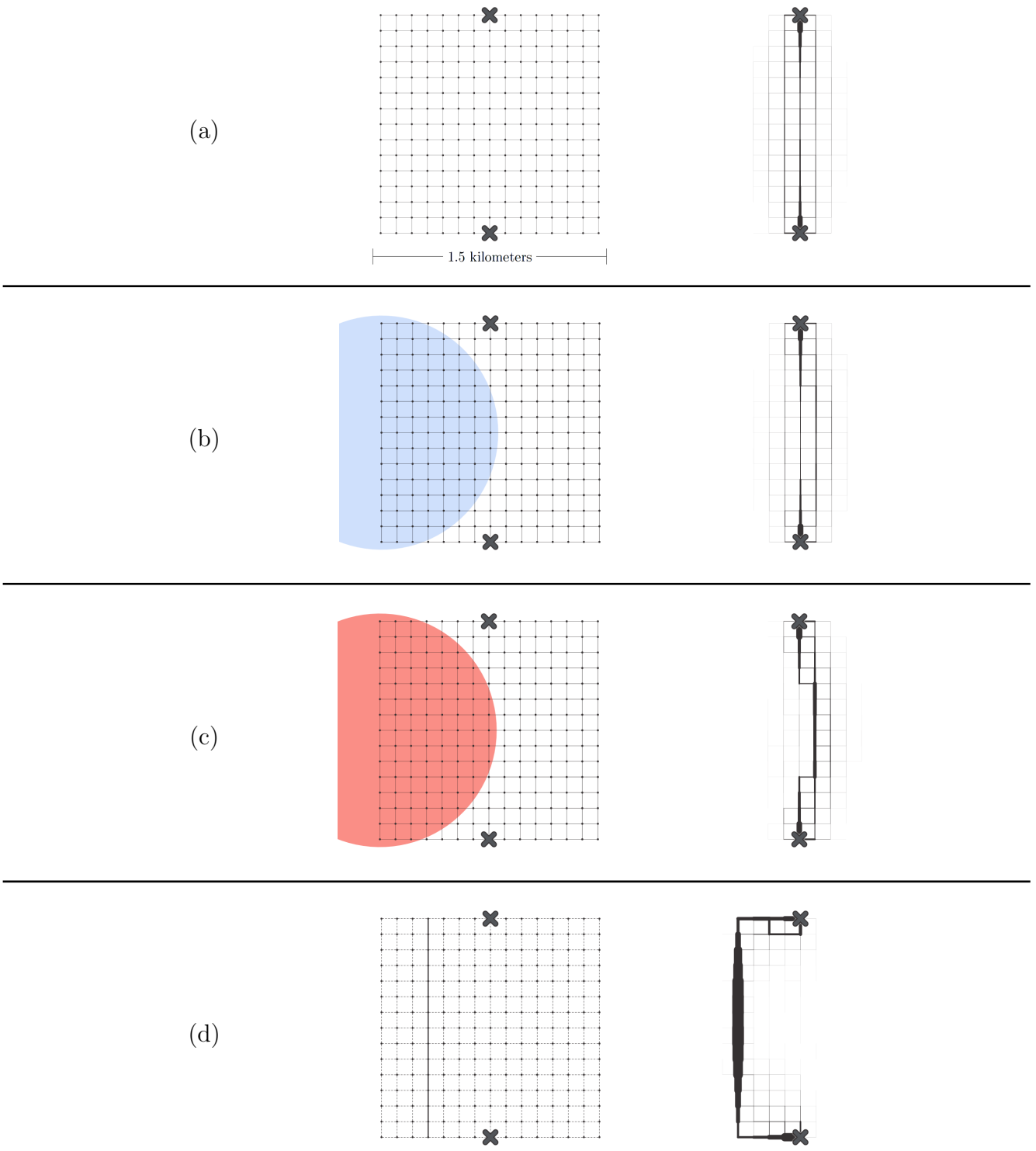

Figure 11: Estimated RPM distribution of walking routes in various scenarios. Each column depicts a hypothetical scenario (upper panel) and 1000 simulated walking routes (lower panel), using point estimates of RPM parameters from Baghdad data. In scenario (a), a walker crosses from bottom to top in a 15×15 square lattice; in the lower panel, thicker edges represent streets that are more likely to be used. In scenario (b), the walker is Sunni and the light blue region is Shia dominated. In scenario (c), the walker is Shia and the dark red region is Sunni dominated. In scenario (d), a vertical line is a major thoroughfare and all other dotted lines represent enclosed residential areas.

tool to examine counterfactual quantities of interest, both by manipulating the walker's preferences (How much more efficient would people be if their aversion could be eliminated?) or altering the network context (How would the same person navigate a different neighborhood? How can urban planners design cities to encourage inter-group contact despite residents' preferences?).

5.5.3 Eliminating Alternative Explanations

In this section, I discuss alternative explanations for these findings and rule them out with additional model specifications. These alternative explanations are ruled out and results are shown to be robust.

One potential concern about the treasure hunt format is that participants might move differently in areas that they know better. If they happen to be more familiar with in-group areas, the difference might distort the estimate of out-group aversion. To address this issue, I add an interaction between `familiarity` with the target (as reported by participants in a debrief) and `direction` toward that target.[22] The basic idea is that people who know the location of their destination are more likely to walk directly toward it. The second panel of figure 9 shows results from this alternative specification. The estimate on this interaction is positive, as expected, but small and insignificant. Other results remain unchanged. Thus, there is no evidence that familiar subjects take more direct routes, and familiarity is not driving the other results. One possible explanation, discussed in section 5.5.4, is that people who are unfamiliar with the target are more likely to ask for directions.

I also consider the possibility that our design did not fully address participants' concerns about safety. In debrief surveys, 16% of participants said that safety was a consideration in their walking routes, despite efforts to choose a safe playing field. If participants avoided out-group areas due to a perceived chance of violence, then their walking patterns would not necessarily indicate taste-based aversion. To test whether this is the case, I interact `safety` concerns with an indicator

[22]The base term, `familiarity` with the endpoint of a treasure-hunt leg, again drops out of the model because it is constant for all edges in a leg.

for out-group areas (`sunni_walker` $\cdot$ `shia_area` + `shia_walker` $\cdot$ `sunni_area`). The third panel of figure 9 shows that, as expected, safety-conscious individuals are much less likely to enter out-group areas ($p = 0.074$). However, results among non-safety-conscious participants are broadly similar: Shia remain highly and significantly averse to Sunni areas, and Sunni are slightly averse to Shia areas (but statistical significance for Sunnis drops to $p = 0.146$).

Finally, I examine whether results change when considering the distance walked through out-group territory, rather than a binary indicator for merely entering it. In the fourth panel of figure 9, two estimates are reported for Sunni walkers—the "baseline" aversion, or the initial discomfort of setting foot into a Shia area, and the "marginal" aversion, or the additional discomfort of each additional 100 meters in Shia territory. The first is insignificant, but estimates show that Sunnis are increasingly reluctant to walk through Shia areas when the distance grows longer. When the two are added together to represent the total discomfort of walking 100 meters across Shia land (a typical distance between intersections), the resulting aversion is comparable to previous estimates and significant at the $p = 0.060$ level. For Shia walkers, both baseline and marginal aversion are significant and in the expected direction.

5.5.4 Discussion

These results strongly suggest that the segregation caused by sectarian conflict is likely to persist beyond the end of this conflict. Aversion is significant even among young and well-educated participants who were primed to explore unfamiliar terrain by the treasure-hunt nature of the task.[23] Among the general population, and during daily life, reluctance to enter out-group-dominated areas is likely to be stronger. It almost certainly translates to an unwillingness to live in these areas. This alone is a sufficient condition for persistent segregation, even if people are still comfortable in mixed neighborhoods (Young, 1998, among others). The intuition behind this literature on "tipping-point" segregation is that when an equally mixed community begins tilting in one di-

[23] In open-ended debrief responses, participants wrote, e.g., "entered new areas", "I saw new things", "it was fun to meet new friends."

rection due to random migration, members of the smaller group will start to flee with increasing urgency. Moreover, segregation tends to be a one-way street, because individuals do not want to be the first to move back into an out-group community.

Heterogeneity in out-group aversion must be interpreted with care. Sectarian identities are not readily manipulable, and their "effects" are arguably undefined (Pearl, 2000; Woodward, 2003). Moreover, our Sunni and Shia participants are not a representative sample of Baghdad residents, and while they are comparable on observable variables, it is possible that unobserved individual- or neighborhood-level variables (e.g., subjects' past exposure to violence or the hostility of neighborhood residents) may be driving this heterogeneity. With these issues in mind, it is interesting to note that Shia participants were more uncomfortable in Sunni areas than vice versa, even though Shia groups have been in power for over a decade. More efficient behavior by Sunnis may reflect a necessary adaptation in a Shia-dominated society. Adaptation would accord with findings from Christia, Knox and Al-Rikabi (n.d.) in a separate study of young Baghdad cafe-goers, where Sunnis are shown to have developed more efficient network strategies for accessing public services: There, among other differences, Sunni were found to be more willing to seek Shia assistance when necessary.[24] Strong aversion among Shia is particularly troubling given their majority status and political power—if Shia are resistant to re-integration, it is unlikely to succeed. Zhang (2004*a*) develops a model showing that segregation remains stable even with one-sided aversion. This is because one group is willing to pay a premium for housing to avoid the other, and the non-averse group moves away in response to price incentives.

Out-group aversion has implications that extend beyond segregation, particularly given Shia control over a wide range of government resources. It also reduces the chances that gradual contact will lead to improved relations. By shaping the way that people roam around their own neighborhoods, aversion in walking routes may also lead to fewer interactions with out-group

[24]Beyond geographic movement and access to public services, we also examine how information is disseminated in sectarian networks in ongoing work. A hypothesis of particular interest is whether Sunnis transmit information more efficiently among themselves, and if so, whether this is because of the tighter-knit structure of Sunni social networks or a behavioral adaptation.

members. In the debrief survey, participants reported whether they asked for directions in each leg. The only significant predictor of "asked for directions" was whether the participant was familiar with their target. In particular, those walking through out-group areas were no less likely to stop and interact with locals—if anything, they were slightly more likely to ask for directions. One possible interpretation is that people are averse to the out-group as a whole, but they are still willing to interact with individual members when required, e.g. at university. If this is the case, walking patterns are reducing the chances for positive contact. However, no data is available on the sect of the person approached, and these results may also be confounded by individual heterogeneity.

6 Future Directions

In this paper, I describe a type of data that is common but underutilized in political science—path data. I show that while dependence between observations presents a statistical challenge, it does not prevent inference on factors shaping the trajectories of paths. The proposed random-path model allows social scientists to assess a wide range of previously untestable hypotheses.

However, much work remains to be done. An `R` package for random-path models is under development and will provide software tools for constructing and cleaning path data, estimating models in a familiar interface, conducting model diagnostics, and visualizing results. In future work, I plan to adapt common statistical procedures for the RPM case. This includes an approach to model selection using likelihood-based cross-validation (van der Laan, Dudoit and Keles, 2004) and sensitivity analyses to test robustness to omitted variables.

Finally, the RPM can be extended to estimate the effects of path-assigned treatments. In ongoing work, I use a model of the assignment process (Rubin, 1991) to simulate a distribution of possible treatment assignments. In the highway case, these can be thought of as proposed highway routes, among which one route is ultimately selected for construction. The resulting distribution

over paths is well-suited for incorporation into Bayesian multilevel models or for drawing causal inferences. For example, this approach allows testing of sharp hypotheses of treatment effects in the Fisherian framework (Rosenbaum, 2002), without the need for untenable assumptions about spatial dependence in the outcome variable. Following Bowers, Fredrickson and Panagopoulos (2013), models of treatment assignment can also be used to evaluate hypotheses about interference between units—a particularly important question in connective infrastructure, where highways can either generate spillover growth in adjacent, unconnected communities or lead to out-migration and decline.

A Appendix

A.1 Exact Enumeration of Paths

Data:
starting node γ_0, terminus γ_k, covariates $\boldsymbol{X}$, parameters β

Result:
$\mathcal{P} \equiv \{\psi : \Omega_\Gamma, |\{\psi\} = |\psi|\}$, set of all paths from γ_0 to γ_k
$\Pr(\Gamma \in \mathcal{P} \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)$, probability that a random walk is a path

Algorithm `PrPath`(γ_0, γ_k, $t = |\gamma|$, $\boldsymbol{X}$, β)
  initialize $\psi = (\gamma_0)$, $t = |\gamma| = 1$, $\mathcal{P} = \{\}$
  populate $\mathcal{P}$ by `recursiveDFS`(γ, t, $\mathcal{N}_{\gamma_{t-1}}$)

  initialize $\Pr(\Gamma \in \mathcal{P} \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta) = 0$
  for $\psi \in \mathcal{P}$ **do**
    $\Pr(\Gamma \in \mathcal{P} \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)$ `+=` $\Pr(\Gamma = \psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)$
  end
  return $\Pr(\Gamma \in \mathcal{P} \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)$

Procedure `recursiveDFS`(ψ, $t = |\psi|$, $\mathcal{N}_{\psi_{t-1}}$)
  for $j \in \mathcal{N}_{\gamma_{t-1}}$ **do**
    if $j = \gamma_k$ **then**
      append j to ψ
      path to terminus found; append ψ to $\mathcal{P}$
    else if $j \in \psi$ **then**
      j already visited; proceed to next neighbor
    else
      append j to ψ
      continue search by `recursiveDFS`(ψ, $t + 1$, $\mathcal{N}_j$)
    end
  end
  pop ψ_t from ψ

Algorithm 2: Calculating the probability that a random walk from γ_0 to γ_k is a path, using depth-first search (DFS) to exhaustively enumerate the set of all paths, $\mathcal{P}$. DFS starts at γ_0 and visits each neighbor in turn, expanding recursively as far as possible until the terminus γ_k is found or no new neighbors are available. The probability that a random walk is in $\mathcal{P}$ is then calculated by summing the probabilities of mutually exclusive events.

A.2 Proof of Proposition 1

This appendix is structured as follows. After introducing the necessary notation, I discuss some properties of the loop-erased random walk. I then outline a procedure that will be used in the proof. Finally, the proof is presented.

A.2.1 Notation

Where the notation in this appendix differs from the simplified exposition in the main text, a note is made.

Let $\tilde{G} = (V, \tilde{E})$ be an undirected, unweighted graph, where $\tilde{E}$ is set of edges (versus an edge-weight matrix in main text). The path $\psi = (V_\psi, E_\psi)$ is a connected subgraph of $\tilde{G}$ that contains no loops or branches (versus a node sequence in main text, where intervening edges were left implicit).

A subgraph of $\tilde{G}$ is a spanning tree if (i) it contains all vertices V, and (ii) every pair of vertices in V is connected by a single unique path on the subgraph. Denote the set of all spanning trees on $\tilde{G}$ as $\mathcal{T}$, and let $\tilde{G}\tau(G)$ be the number of such trees. The path ψ is "on" a particular spanning tree $T = (V, E_T)$ if it is a subgraph of T; this holds if $E_\psi \subseteq E_T$, since necessarily $V_\psi \subseteq V$.

A.2.2 LERW Properties

Wilson's algorithm (Wilson, 1996) takes as input the graph $\tilde{G}$ and some ordering of its nodes $U = (u_1, \cdots, u_N)$, then returns a random sample from the set of possible spanning trees, $\mathcal{T}$. In brief, the algorithm starts from u_2 and performs a LERW until u_1 is reached, then marks all nodes and edges along the resulting path as visited. It then proceeds to the next node in U that has not been previously visited, performs a LERW until reaching any previously visited node, and again marks everything along that path as "previously visited." This process is iterated until all nodes have beeen visited. The resulting set of visited nodes and edges is a spanning tree on $\tilde{G}$; Wilson showed that for any choice of U, the procedure samples each element of $\mathcal{T}$ with equal probability.

See also Lawler (1999)[pp. 211–212] for a more illuminating proof.

Corollary A.1. $\Pr\left(\textit{LERW}(\tilde{\mathrm{G}}, \mathrm{u}_2, \mathrm{u}_1) = \psi\right)$ *is proportional to the number of spanning trees on* $\tilde{G}$ *that contain* ψ*.*

Proof. Let W be a spanning-tree-valued random variable whose probability mass is uniformly distributed over elements of $\mathcal{T}$. Wilson's algorithm is a procedure to sample W, in which $\texttt{LERW}(\tilde{G}, u_2, u_1)$ is the first step. A spanning tree contains one unique path between u_2 and u_1. Therefore,

$$\Pr\left(W = T, \texttt{LERW}(\tilde{G}, u_2, u_1) = \gamma\right) = \begin{cases} \Pr(W = T) & \text{if} \quad \gamma \text{ is on } T \\ 0 & \text{if} \quad \gamma \text{ is not on } T \end{cases}$$

It immediately follows that

$$\begin{aligned} \Pr\left(\texttt{LERW}(\tilde{G}, u_2, u_1) = \psi\right) &= \sum_{T \in \mathcal{T}} \Pr\left(\texttt{LERW}(\tilde{G}, u_2, u_1) = \psi \mid W = T\right) \Pr(W = T) \\ &= \sum_{T \in \mathcal{T}} \mathbf{1}\{\psi \text{ is a subgraph of } T\} \frac{1}{|\mathcal{T}|}. \end{aligned} \tag{3}$$

□

A.2.3 Deletion-Contraction Recurrence

I now outline a method that will be needed to count trees that contain a path. Consider an arbitrary edge, e, in $\tilde{G}$. The deletion-contraction recurrence (see, e.g. Bollobás, 1998, Theorem X.5.10, pp. 351–353) states that $\mathcal{T}$ can be divided into two disjoint sets: the set of spanning trees that do not use e, and the set of spanning trees that do. The former is in one-to-one correspondence with the set of spanning trees on the *deletion*, denoted $\tilde{G} - e$, formed by cutting e. The latter is similarly in one-to-one correspondence with the set of spanning trees on the *contraction*, $\tilde{G}/e$, formed by fusing the endpoints of e into a single node.[25] Thus, $\tau(\tilde{G}) = \tau(\tilde{G} - e) + \tau(\tilde{G}/e)$.

[25] Note that this procedure may result in a multigraph.

A.2.4 Proof

We are now ready to prove Proposition 1.

By recursive deletion-contraction, there is a bijection between (i) the set of spanning trees on $\tilde{G}$ that contain ψ as a subgraph and (ii) the set of spanning trees on the iterated contraction $\tilde{G}/e_{\psi,1}/\cdots/e_{\psi,K}$, where $e_{\psi,t}$ is the t-th edge in ψ. Kirchoff's matrix-tree theorem states that the number of spanning trees on a graph is given by the determinant of any minor of the graph's Laplacian matrix,

$$\tau(\tilde{G}) = \det\ L_{(-i,-j)}(\tilde{G}),$$

for any i and j, where the Laplacian, $L(\tilde{G}) = \tilde{D} - \tilde{A}$, is the diagonal degree matrix less the adjacency matrix. Substituting into equation 3 yields

$$f_{\text{LERW}}(\psi) = \Pr(\texttt{LERW}(\tilde{G}, u_2, u_1) = \psi) = \frac{1}{\tau(\tilde{G})}\det\ L_{(-i,-j)}(\tilde{G}/e_{\psi,1}/\cdots/e_{\psi,K}),$$

for the LERW importance-sampling distribution, versus the target uniform distribution $f(\psi) = \frac{1}{|\mathcal{P}|}$. The corrective weight for importance sampling is the ratio of the latter relative to the former, which is

$$\frac{\tau(\tilde{G})}{|\mathcal{P}|\ \det\ L_{(-i,-j)}(\tilde{G}/e_{\psi,1}/\cdots/e_{\psi,K})}$$

□

A.3 Details of RPM Estimation by MCMC

In this appendix, I discuss computational issues in estimation and provide an algorithm for sampling from the RPM posterior via MCMC. First, the sampling-reweighting procedure involves large numbers of simulated paths and expensive matrix determinants. Rather than repeating algorithm 1 in its entirety for each MH proposal, it is clearly advantageous to pre-compute a single batch of paths and their weights. This has the ancillary benefit of reducing noise in the MH ac-

ceptance ratio, as the simulated likelihood of both current and proposed parameters are estimated with the same path-set.

To evaluate the likelihood at any point in the parameter space, algorithm 1 must compute the unconditional (random-walk) probabilities of many paths. Because MCMC methods frequently revisit a relatively small, dense-probability region in the parameter space, a naïve implementation will spend considerable time repeatedly evaluating the likelihood at infinitesimally differing points. An alternative that considerably reduces running time, at the expense of initialization time and memory, is to pre-compute a finely gridded piecewise-constant approximation of the likelihood across a wide subspace. However, this contradicts the spirit of MCMC and is computationally infeasible for parameter spaces of moderate dimension. I implement a compromise by lazy evaluation of the likelihood over the parameter grid. In areas that are never sampled by MH, the computational cost is never incurred and memory usage is greatly decreased. After a cell is sampled by the MH proposal distribution, the likelihood is evaluated and cached for future use, or "memoized." Thus, chains will accelerate as they grow longer or more numerous, particularly when sampling the high-posterior-density region.

Data:

starting node γ_0, teminus γ_k, covariates $\boldsymbol{X}$

unweighted graph $\tilde{G}$, number of path simulations S

initial parameters $\beta^{(0)}$, gridded parameter space $\tilde{\mathcal{B}}$

number of Metropolis-Hastings samples R, proposal distribution $Q(\beta^*; \beta^{(t)})$

Result:

R correlated samples from posterior of parameters β

Algorithm `ChainMH`(γ, $\boldsymbol{X}$, $\beta^{(0)}$, $\tilde{\mathcal{B}}$, Q)

for $s \in 1, \cdots, S$ **do**

draw $\psi_s \sim \texttt{LERW}(\tilde{G}, \gamma_0, \gamma_k)$

calculate $w_s = \frac{1}{\det\ L_{(-i,-j)}(\tilde{G}/\psi_s)}$

end

set evaluated$_{\tilde{\beta}}$ = `FALSE` for all $\tilde{\beta} \in \tilde{B}$

for $r \in 0, \cdots, R$ **do**

draw proposed parameters $\beta^* \sim Q(\beta^*; \beta^{(r)})$

if parameter space is discretized **then**

calculate acceptance ratio $\alpha = \frac{\texttt{ApproxSimLikelihood}(\beta^*)}{\texttt{ApproxSimLikelihood}(\beta^{(r)})}$

else

calculate acceptance ratio $\alpha = \frac{\frac{\Pr(\Gamma=\gamma | v_0=\gamma_0, v_K=\gamma_k, \boldsymbol{X}, \beta^*)}{\sum_{l=1}^{s} w_s \Pr(\Gamma=\psi_s | v_0=\gamma_0, v_K=\gamma_k, \boldsymbol{X}, \beta^*)}}{\frac{\Pr\left(\Gamma=\gamma | v_0=\gamma_0, v_K=\gamma_k, \boldsymbol{X}, \beta^{(r)}\right)}{\sum_{l=1}^{s} w_s \Pr\left(\Gamma=\psi_s | v_0=\gamma_0, v_K=\gamma_k, \boldsymbol{X}, \beta^{(r)}\right)}}$

end

if $\alpha < 1$ and jump $\sim \text{Bern}(\alpha)$ **then**

set $\beta^{(r+1)} = \beta^*$

else

set $\beta^{(r+1)} = \beta^{(r)}$

end

end

return $\beta^{(0)}, \cdots, \beta^{(R)}$

Procedure `ApproxSimLikelihood`(β)

set $\tilde{\beta}$ to center of grid cell in $\tilde{B}$ containing β

if evaluated$_{\tilde{\beta}}$ **then**

return precomputed $\hat{\mathcal{L}}(\tilde{\beta} \mid \boldsymbol{X}, \gamma)$

else

set evaluated$_{\tilde{\beta}}$ = `TRUE`

return and cache $\hat{\mathcal{L}}(\tilde{\beta} \mid \boldsymbol{X}, \gamma) \ = \ \frac{\Pr\left(\Gamma=\gamma | v_0=\gamma_0, v_K=\gamma_k, \boldsymbol{X}, \tilde{\beta}\right)}{\sum_{l=1}^{s} w_s \Pr\left(\Gamma=\psi_s | v_0=\gamma_0, v_K=\gamma_k, \boldsymbol{X}, \tilde{\beta}\right)}$

end

Algorithm 3: Implementing Metropolis-Hastings for a random-path model. Simulated likelihood calculations are memoized so that chains accelerate as they sample the highest-posterior-density region.

B Simulation

In this section, I first demonstrate the properties of the random-path distribution with a naturalistic simulation. I then conduct a validation test in which a single path is drawn and its parameters are estimated by MCMC. This procedure is repeated at various sample sizes and graph resolutions in order to assess the consistency of the estimation procedure.

B.1 Simulation Distribution of Random Paths

The simulation ground is a virtual Hawai'i Island, rasterized into square cells of varying size. Each cell is connected to a tic-tac-toe board consisting of the 8 adjacent cells and excluding self-loops. I assume that a single road will be constructed from the western economic center, Kona, to the county seat in the east, Hilo. Figure 12 depicts the difference between a typical random walk and path on the unweighted graph; it is perhaps unnecessary to point out that the random path bears a closer resemblance to actual Hawai'ian state highways.

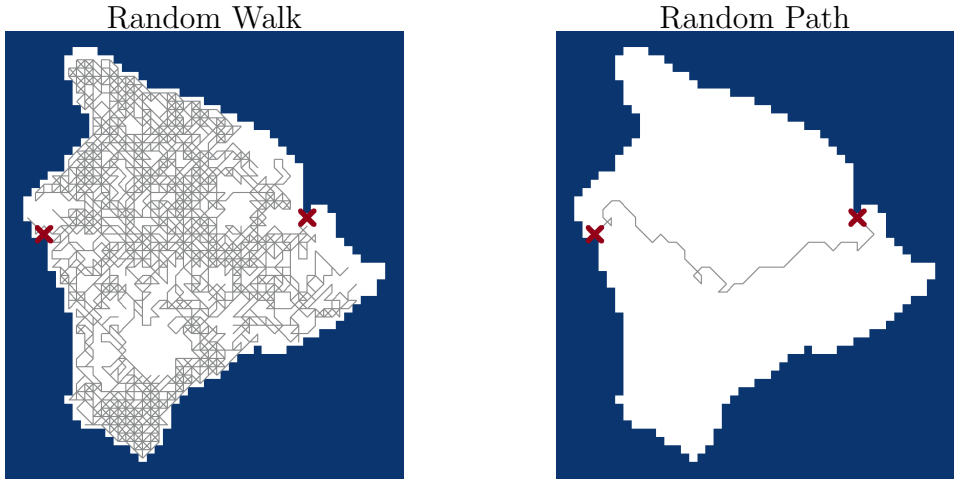

Figure 12: One draw each from the random walk and random path distributions, on a 50×50 grid, with all parameters set to zero.

One might reasonably expect a Hawai'ian road to avoid excessively mountainous regions, while passing through as many villages as possible without deviating too far from a direct course. As a

point of reference, actual state highways on the Big Island are roughly Θ-shaped, consisting of a circular coastal highway and Saddle Road, which cuts directly from Kona to Hilo. To capture this behavior, I include transformations of three covariates: (1) directness $\texttt{dir}_{ij}$, or how much closer the $i \to j$ step brings a walker to the target; (2) elevation $\texttt{elev}_j$; and (3) population "gravity."[26] The covariates are shown in figure 13. For a walker at cell i, the unconditional (random-walk) probability of stepping to adjacent cell j is

$$\frac{\exp\left(\beta_{\text{dir}} \cdot \texttt{dir}_{ij} + \beta_{\text{elev}}\texttt{elev}_j + \beta_{\text{pop}} \cdot \texttt{lpop}_{ij}\right)}{\sum_{j' \in N_i} \exp\left(\beta_{\text{dir}} \cdot \texttt{dir}_{ij'} + \beta_{\text{elev}}\texttt{elev}_{j'} + \beta_{\text{pop}} \cdot \texttt{pop}_{ij'}\right)}.$$

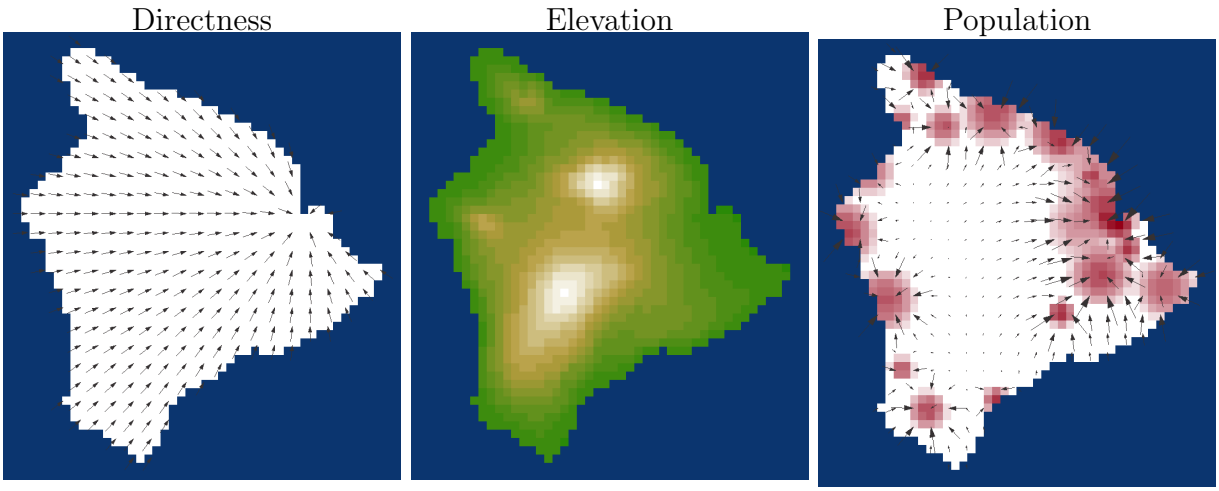

Figure 13: RPM covariates on a 50×50 grid. Direction toward target (left) indicated by arrows. Elevation (center) in terrain colors, green at sea level and white at ∼4,000 m, around the peaks of Mauna Kea and Mauna Loa. Log-population (right) plotted in red, with higher density in more opaque regions. Arrows show the direction of population gravitational pull, with arrow size indicating force.

The random-path distribution is the conditional random-walk distribution, given that the walk does not contain cycles. The simulation distribution of a random-path model is the result

[26] Directness is calculated as the inner product of the step vector $(\texttt{location}_j - \texttt{location}_i)$ with a unit vector pointing from i to Hilo. Elevation is rasterized by averaging National Elevation Dataset values within j, then scaled and exponentiated to increase separation. Raster-cell population is generated to be consistent with 1940 census tract data (with Gaussian allocation of tract population around approximate coordinates of in-tract villages); each cell is assumed to generate a gravitational pull proportional to its log-population and the inverse squared distance, and $\texttt{pop}_{ij}$ is operationalized as the inner product of the step-vector $i \to j$ with the aggregate gravitational field at i.

of importance-sampling S paths, calculating the random-walk probability of each, then resampling from the S paths with probability proportional to random-walk probability times inverse-importance weights. The simulation distribution converges to the true RPM distribution as S increases; in the illustrations that follow, I use $S = 10^6$ and resample 10^2 paths.

In figure 14, I show the result of increasing β_{dir}. The left panel in figure 14 is a larger sample from the baseline distribution with all parameters set to zero (the same RPM that generated the right panel of figure 12). The baseline distribution is the path-conditioned version of a random walk in which all adjacent cells are equally likely. After conditioning to walks that contain no cycles, the random-path distribution exhibits a strong baseline preference for shorter (more direct) paths. This is because the longer a walk continues, the more likely it is to double back on itself. In the right panel, I show that this natural tendency can be reinforced by increasing β_{dir}; at higher values, the random path distribution becomes tighter and more focused. Figures 15 and 16 depict the effects of β_{elev} and β_{pop}, respectively.

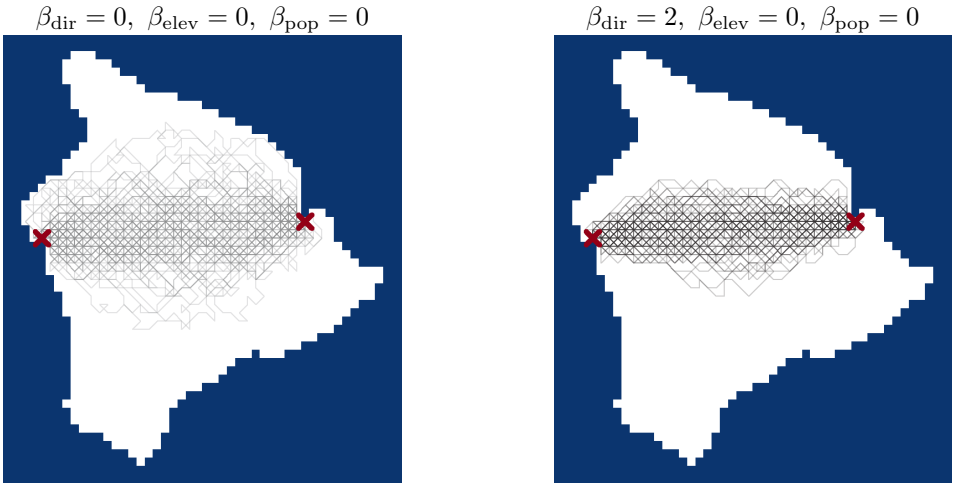

Figure 14: Higher values of β_{dir} (right) result in a tighter distribution with more direct paths than the baseline (left).

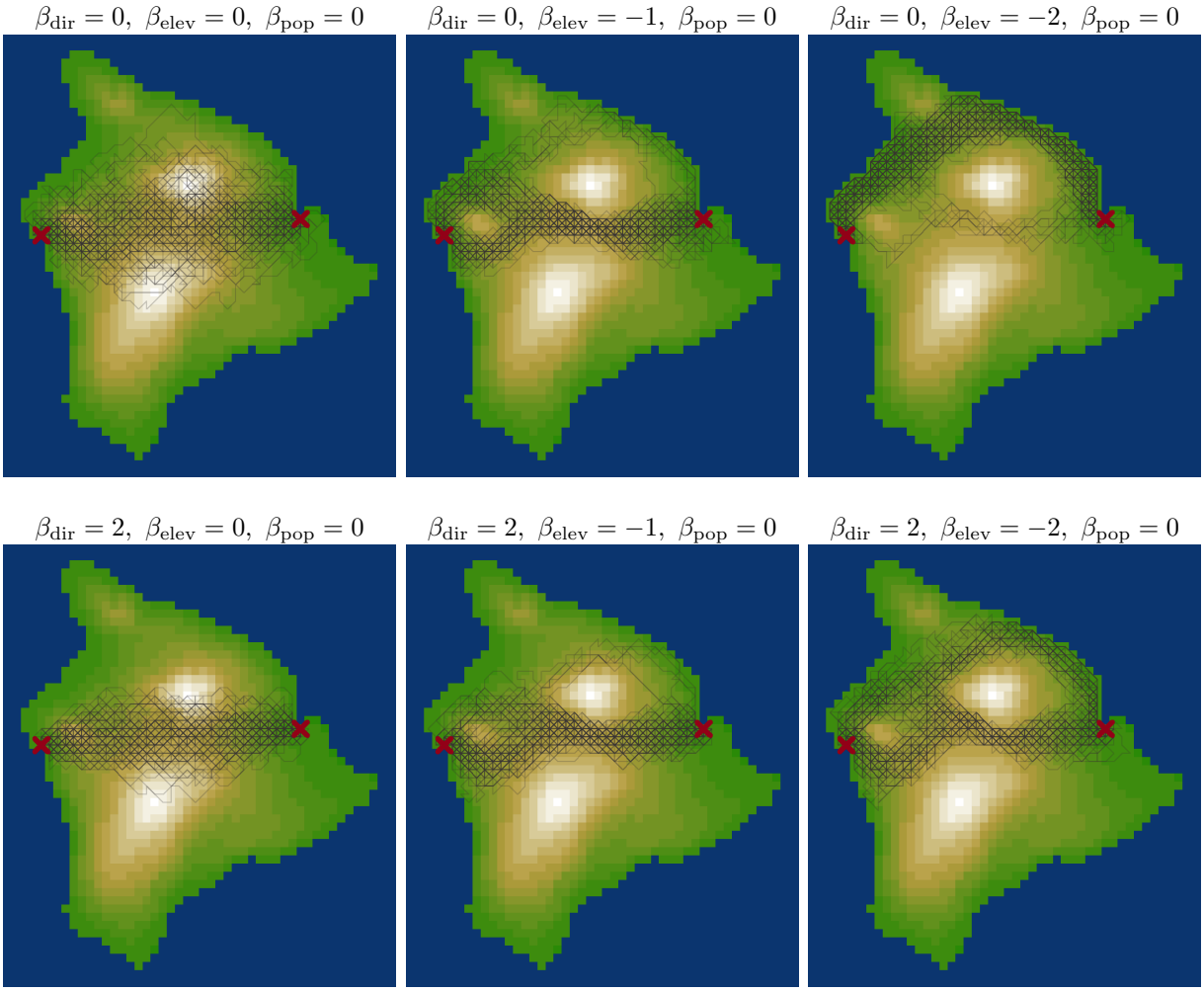

Figure 15: Increasingly negative values of β_{elev} (moving right) result in distributions that avoid mountainous regions. However, this tendency can be partially overcome by higher values of β_{dir} (lower plots), which drive the path distribution over the saddle pass directly toward Hilo.

B.2 Validating the Estimation Procedure

B.2.1 Convergence

I first assess the MCMC convergence of the RPM posterior by randomly drawing a single path from RPM($\beta_{\text{dir}} = 0, \beta_{\text{elev}} = -1, \beta_{\text{pop}} = 0.5$). The true distribution was chosen such that with a single draw, equivalent to perusing a map, a reasonable human observer would consider $\hat{\beta}_{\text{elev}}$ to be negative and statistically significant and both $\hat{\beta}_{\text{dir}}$ and $\hat{\beta}_{\text{pop}}$ to be perhaps slightly positive but indistinguishable from zero. In fact, the first sampled path (shown in figure 17) captures this

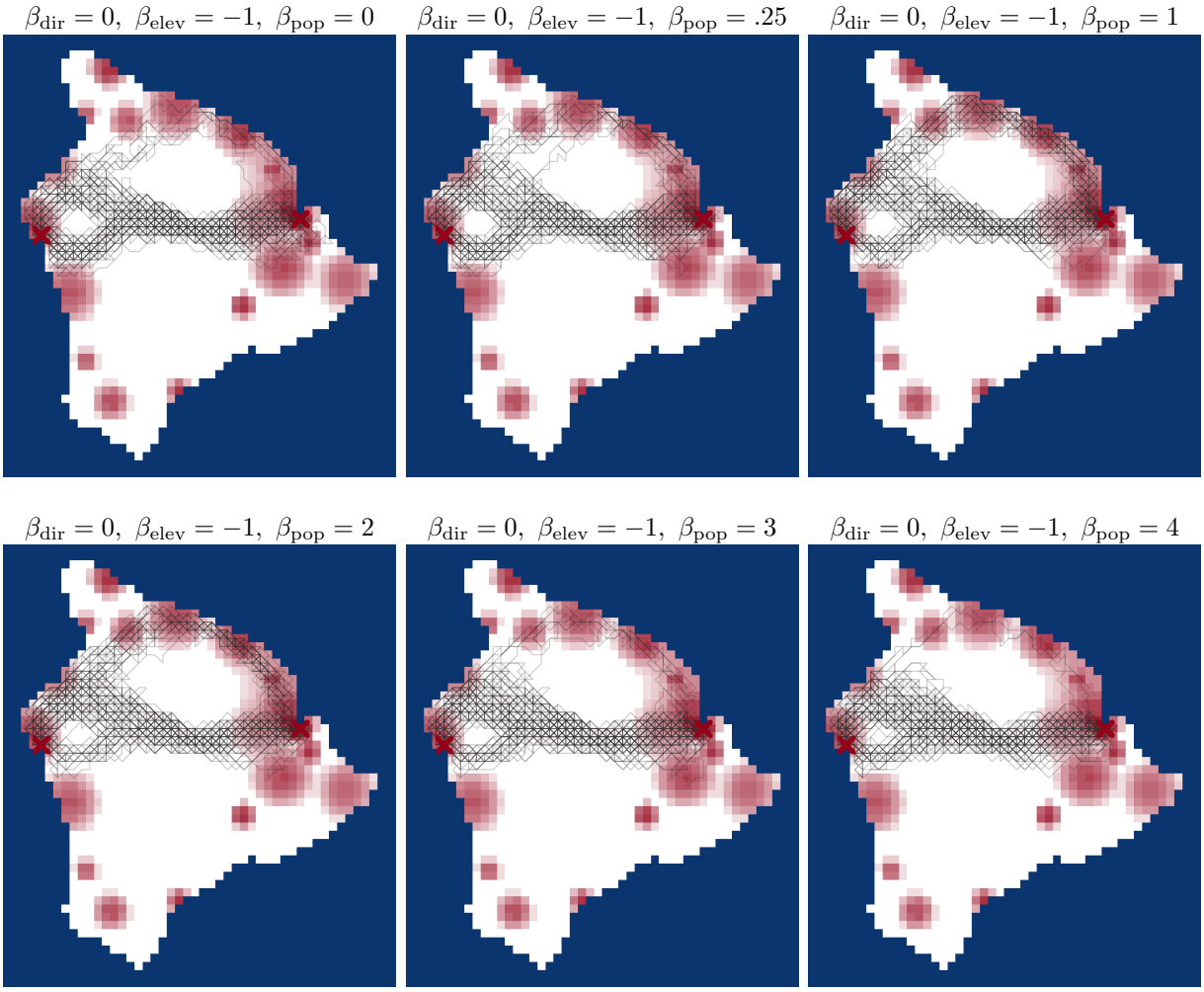

Figure 16: Small increases in β_{pop} (upper row, moving right) make coastal paths more likely to visit small towns instead of passing by (esp. Waimea, on the northern peninsula), then begin to redirect paths away from the saddle pass and toward coastal population centers. At very large values, however, this effect reverses as paths are pulled directly over the pass by the strong gravitational pull of the large Hilo population (lower right).

intent nicely. Starting in the west at Kona, the path tracks the city limits as it diverts around Hualalai, the volcano just outside the city, then traverses the saddle pass before exiting with a slight flourish. I examine the extent to which the RPM posterior reflects these patterns. The effective number of observations in a single path, after accounting for dependence, is somewhere in $[1, k]$.

I evaluate the mixing of MH-sampled MCMC and the resulting estimates. Chain length was

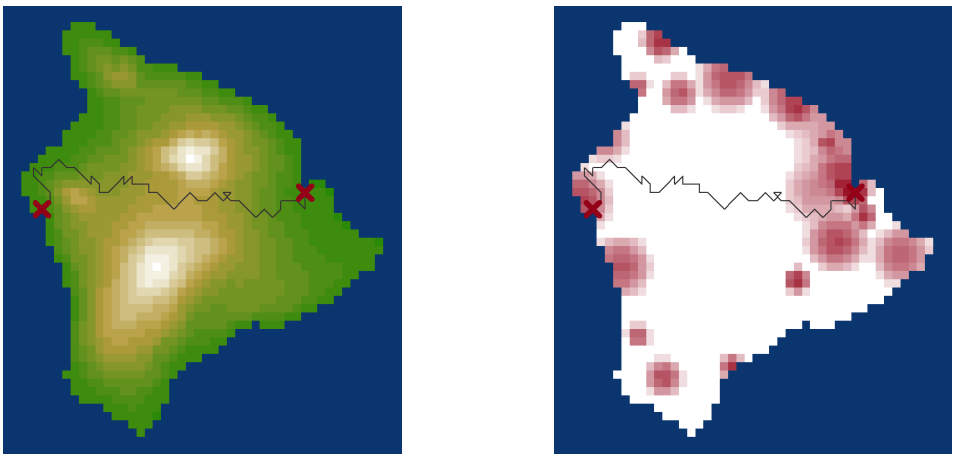

Figure 17: A single draw from RPM($\beta_{\text{dir}} = 0, \beta_{\text{elev}} = -1, \beta_{\text{pop}} = 0.5$), plotted against elevation (left) and population (right).

5,000 draws and the reduction in effective posterior sample size due to autocorrelation was a factor of roughly 15, differing only slightly by parameter. This left an effective sample size of roughly 300–350 and sampling standard errors of parameter posterior means between 0.02 and 0.05—more than an order of magnitude smaller than estimated posterior standard deviations, and quite acceptable for present purposes. Chains for each parameter, posterior means, and 95% posterior credible intervals are shown in figure 18; elevation was estimated to be negative and correctly signed, while all other parameter estimates were insignificant.

B.2.2 Consistency

Next, I examine the Bayesian consistency of the RPM estimation procedure. Specifically, I generate paths according to a true distribution, then evaluate whether the posterior means and variances of the distribution parameters go to the true parameter and zero, respectively, (*i*) as the number of paths increase, but approximate length of each path remains fixed; and (*ii*) as paths grow longer, but the number of paths remain fixed. To test (*i*), I examine the RPM posterior distribution given sample sizes of 1, 2, 4, 8, and 16 paths between fixed endpoints on the same graph. For (*ii*), I re-rasterize the Hawai'i simulation ground into 10×10, 20×20, and 40×40 square grids, then compare

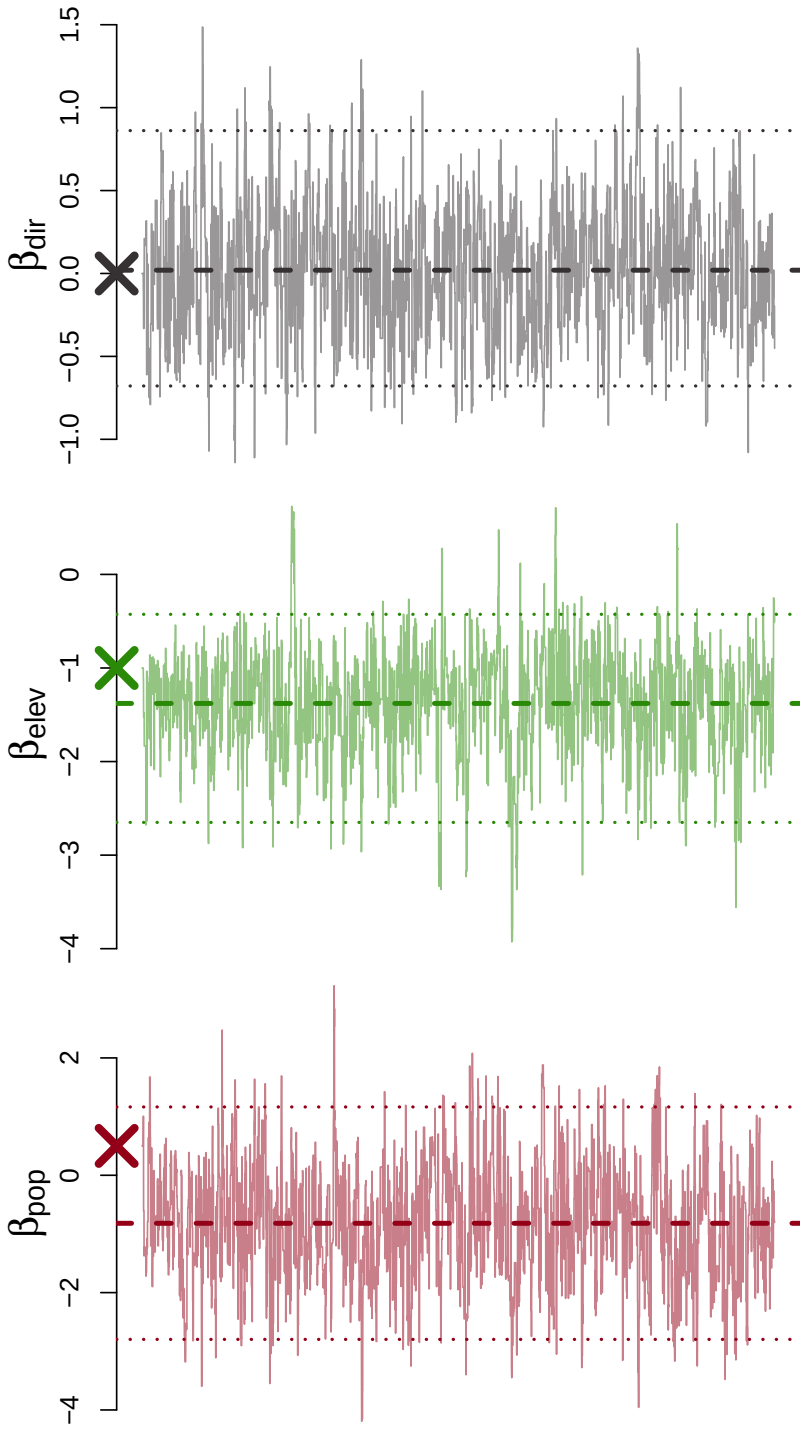

Figure 18: RPM posterior for the path depicted in figure 17, with sampled parameter values on the vertical axis and iterations on the horizontal. Parameter posterior means (dashed) and 95% marginal posterior credible intervals (dotted) are plotted horizontally over each chain. The true parameter is marked with a "×" on the vertical axis.

the posterior on these graphs given a fixed number of sampled paths. Given computational constraints, I focus here on the elevation parameter only. The true parameter used below is $\beta_{\text{elev}} = -2$.

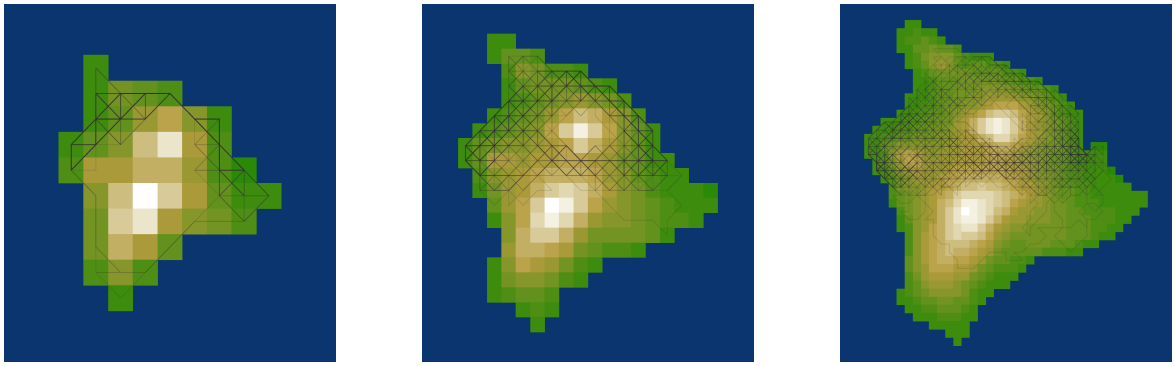

Figure 19: 100 draws from RPM($\beta_{\text{elev}} = -2$) on 10×10, 20×20, and 40×40 square grids

The true RPM distributions are shown in figure 19 for each grid size. The procedure used is as follows: For the 10×10 grid, a single path was sampled and its posterior distribution was approximated by algorithm 3; this corresponds to the first horizontal line in the top-left panel of figure 20. In total, 100 single paths were drawn on the 10×10 grid—results are shown in the top-left panel.

Next, paths were sampled from the true model, two at a time; the approximate posteriors for 100 sampled pairs are shown in the second panel in the top row. This was repeated with samples of 4, 8, and 16 paths. The entire process was repeated for the 20×20 and 40×40 grids (second and third row of panels).

In this simulation, results initially show that estimates are correctly signed but unmistakably biased toward zero for short paths. Variance goes to zero, but bias does not disappear as the number of short paths increase. This suggests that for small graphs, a bias-correction step, such as simulating paths from the posterior and re-estimating, may be necessary. As the graph grows larger and paths grow longer, this bias disappears and estimates converge toward the true parameter.

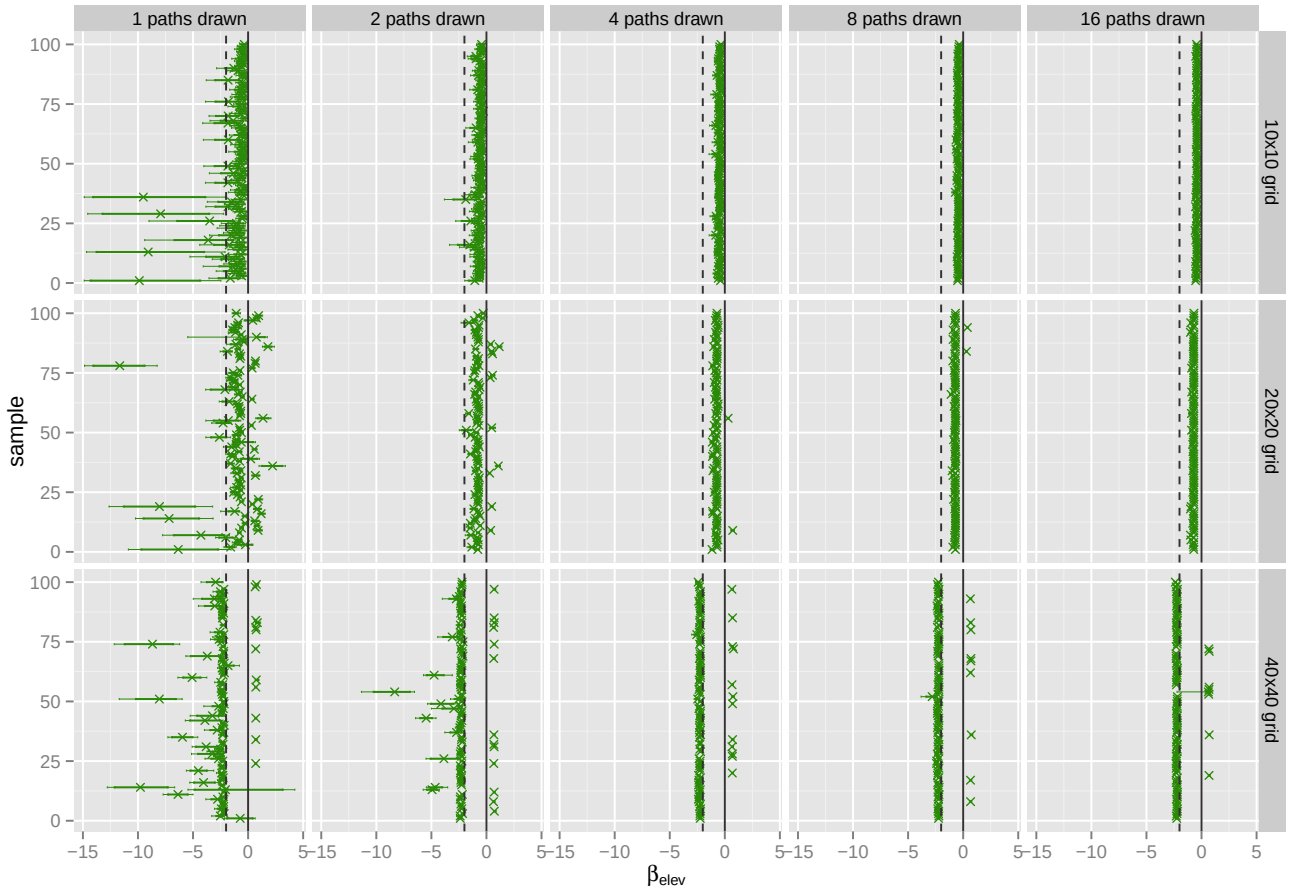

Figure 20: For each combination of sample size and grid size, 100 samples were drawn. Posterior means are marked with "×", 95% credible intervals with thin horizontal green lines, and 80% intervals with thick green horizontal lines. The true parameter is shown with a vertical black dotted line. Results show that estimator variance converges to zero as sample size increases, but some bias remains when paths are short. This bias disappears as paths grow longer.

C Convergence of U.S. Interstate Highways Estimates

The posterior of RPM parameters in the U.S. Interstate Highway application was simulated by MCMC. Five chains, of 10,000 samples each, were initialized at overdispersed locations. After a burn-in of 2,000 iterations, visual diagnostics show excellent mixing and low autocorrelation relative to chain length.

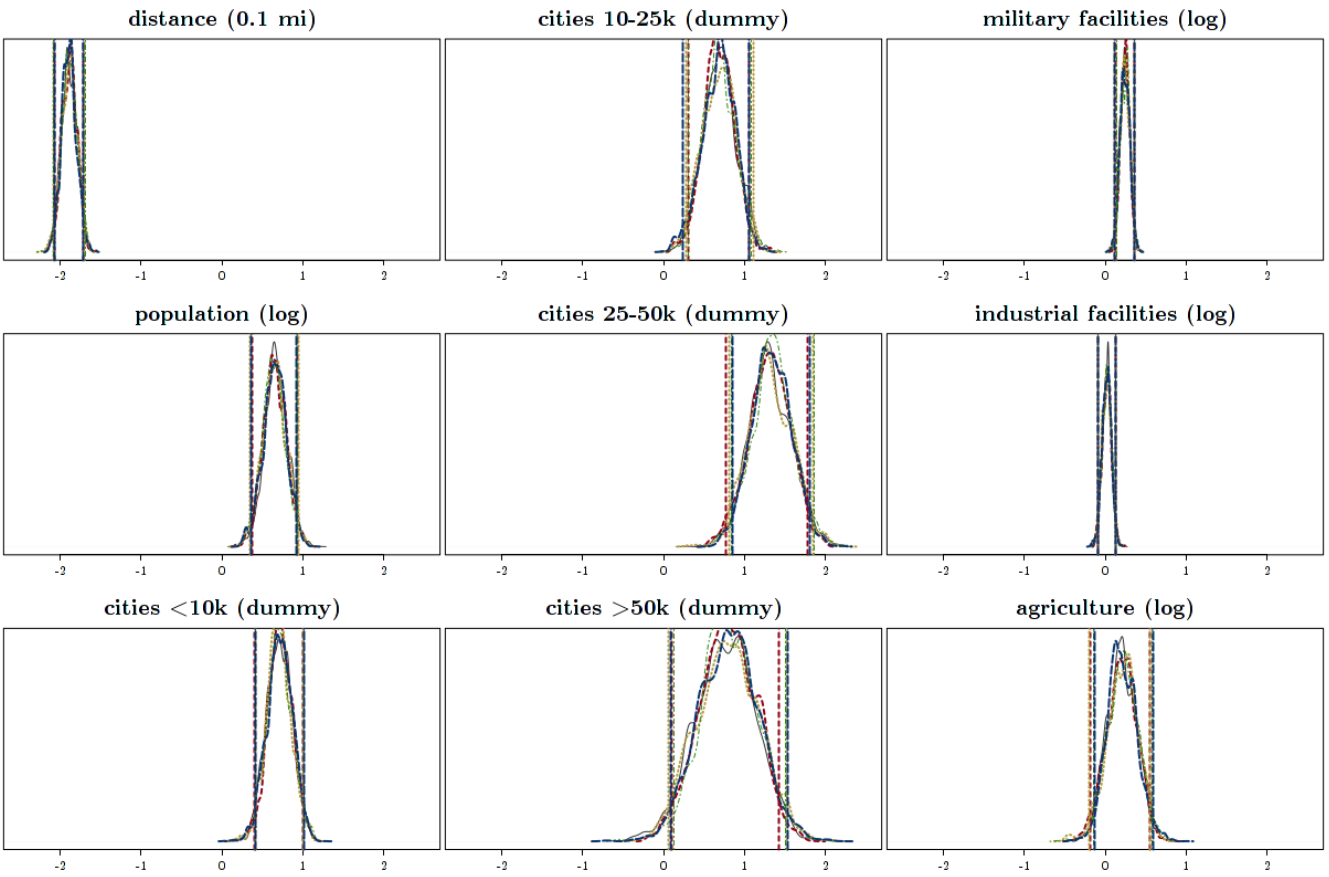

Figure 21: After discarding the first 2,000 iterations of each chain as burn-in, marginal posterior densities of RPM parameters over the remaining 8,000 iterations are extremely similar. Separate colors and line types represent each chain. Vertical bars represent 2.5 and 97.5-th posterior percentiles.

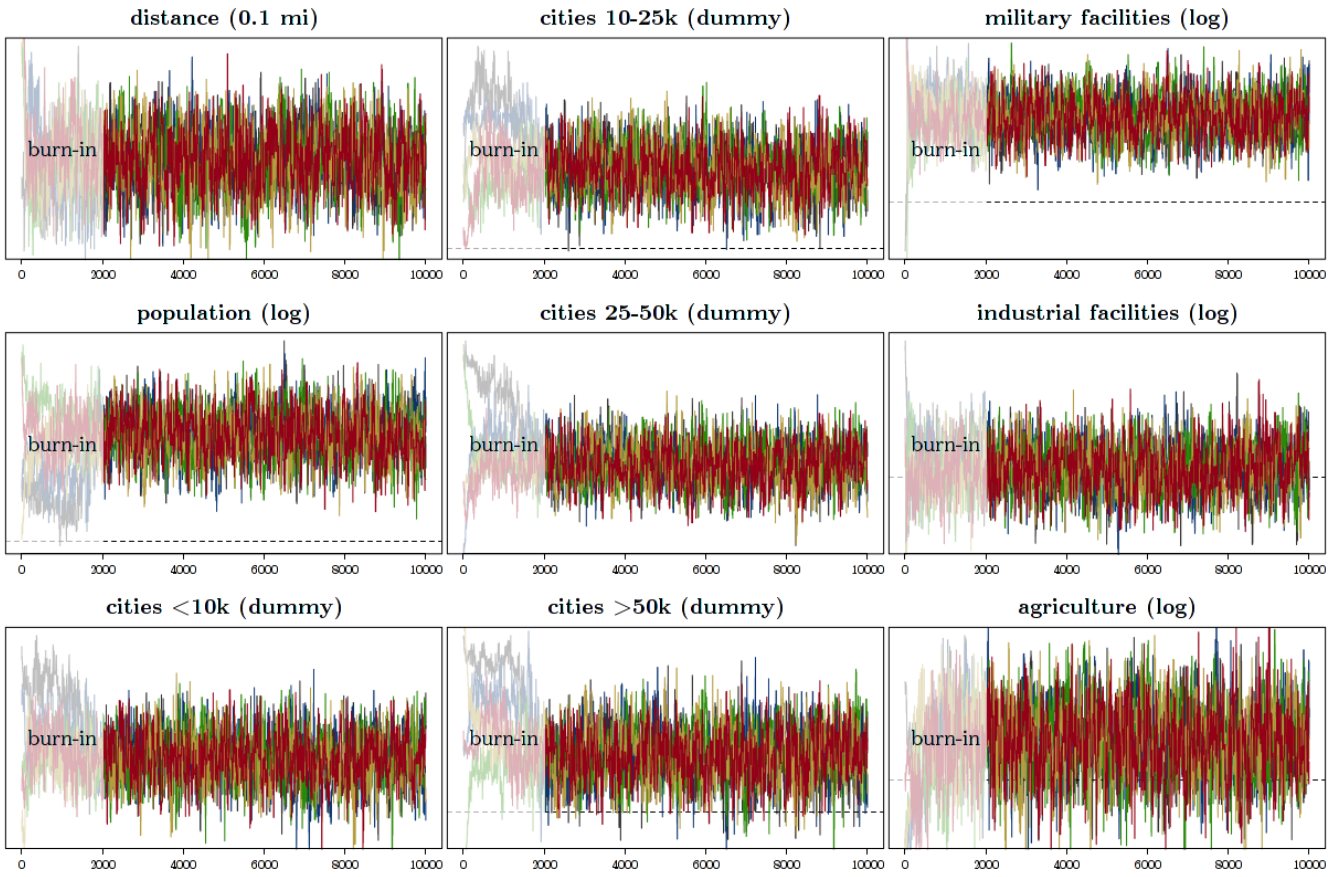

Figure 22: Traces of five chains, denoted by color. Visual inspection suggests that a burn-in of 2,000 iterations is adequate and that autocorrelation is low relative to chain length.

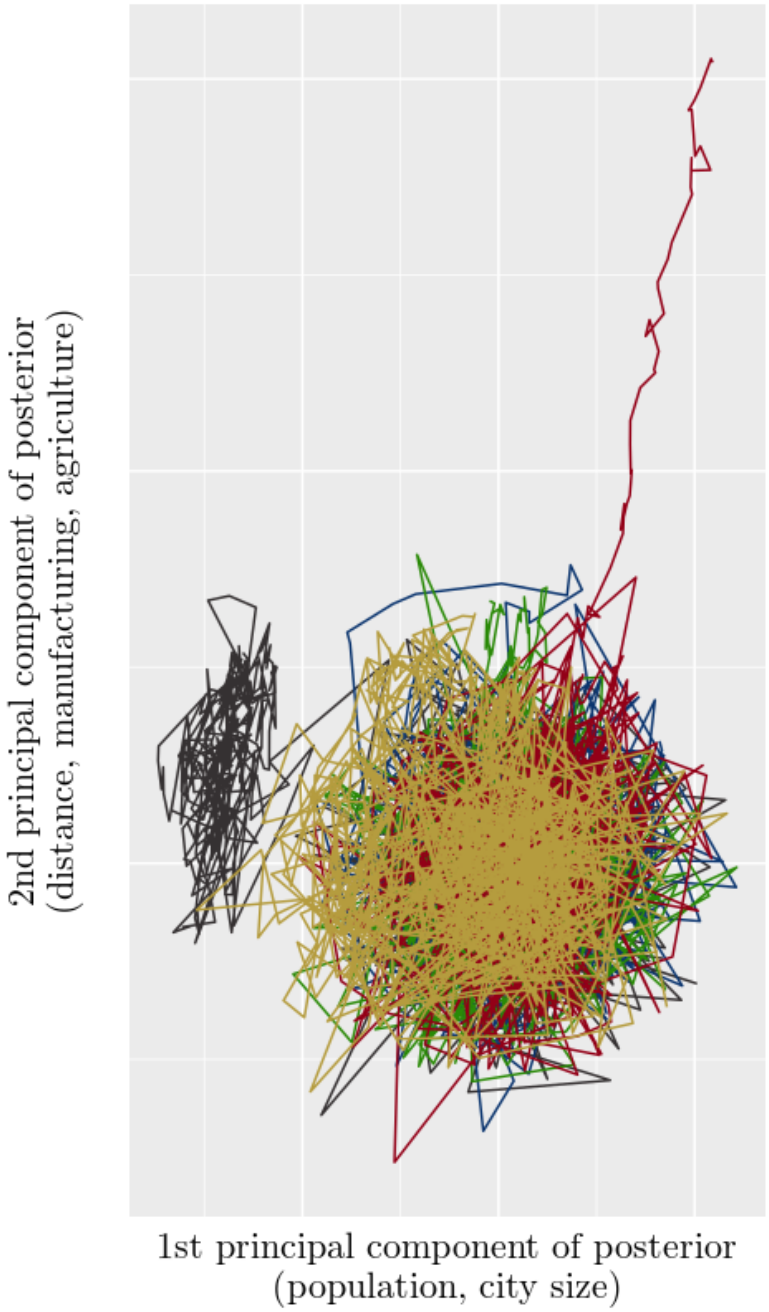

Figure 23: Visualizing MCMC chains with the first two principal components of the posterior distribution. The plot shows that chains initialized at overdispersed starting positions converge to the same region in the parameter space, with excellent mixing. The first component roughly captures population-related covariates, and the second is a mix of the remaining covariates.

D Convergence of Baghdad Sectarianism Estimates

The posterior of RPM parameters in the Baghdad walks application was simulated by MCMC. Three chains, of 10,000 samples each, were initialized at overdispersed locations. After a burn-in of 4,000 iterations, visual diagnostics show excellent mixing and low autocorrelation relative to chain length.

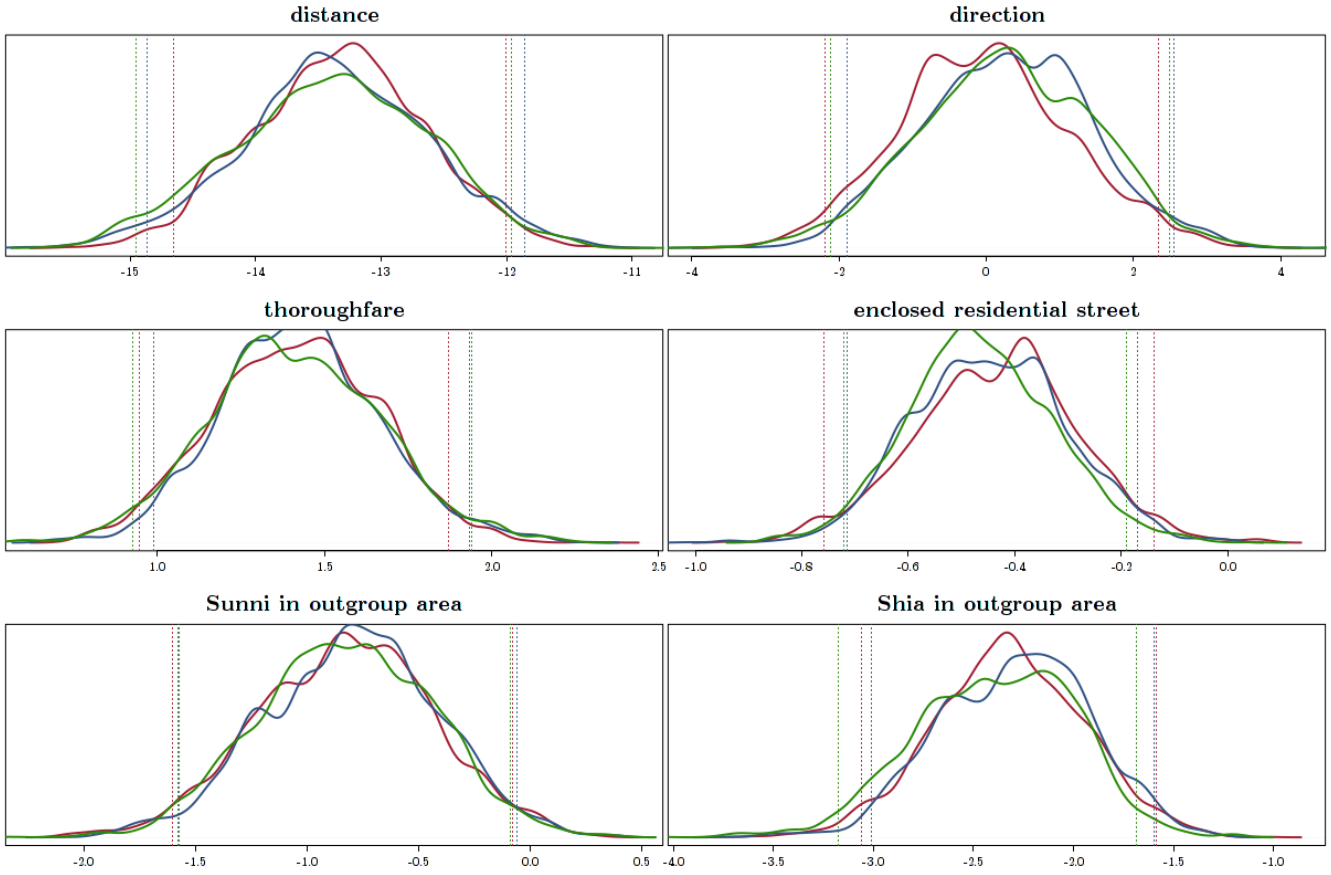

Figure 24: After discarding the first 4,000 iterations of each chain as burn-in, marginal posterior densities of RPM parameters over the remaining 6,000 iterations are extremely similar. Separate colors represent each chain. Vertical bars represent 2.5 and 97.5-th posterior percentiles.

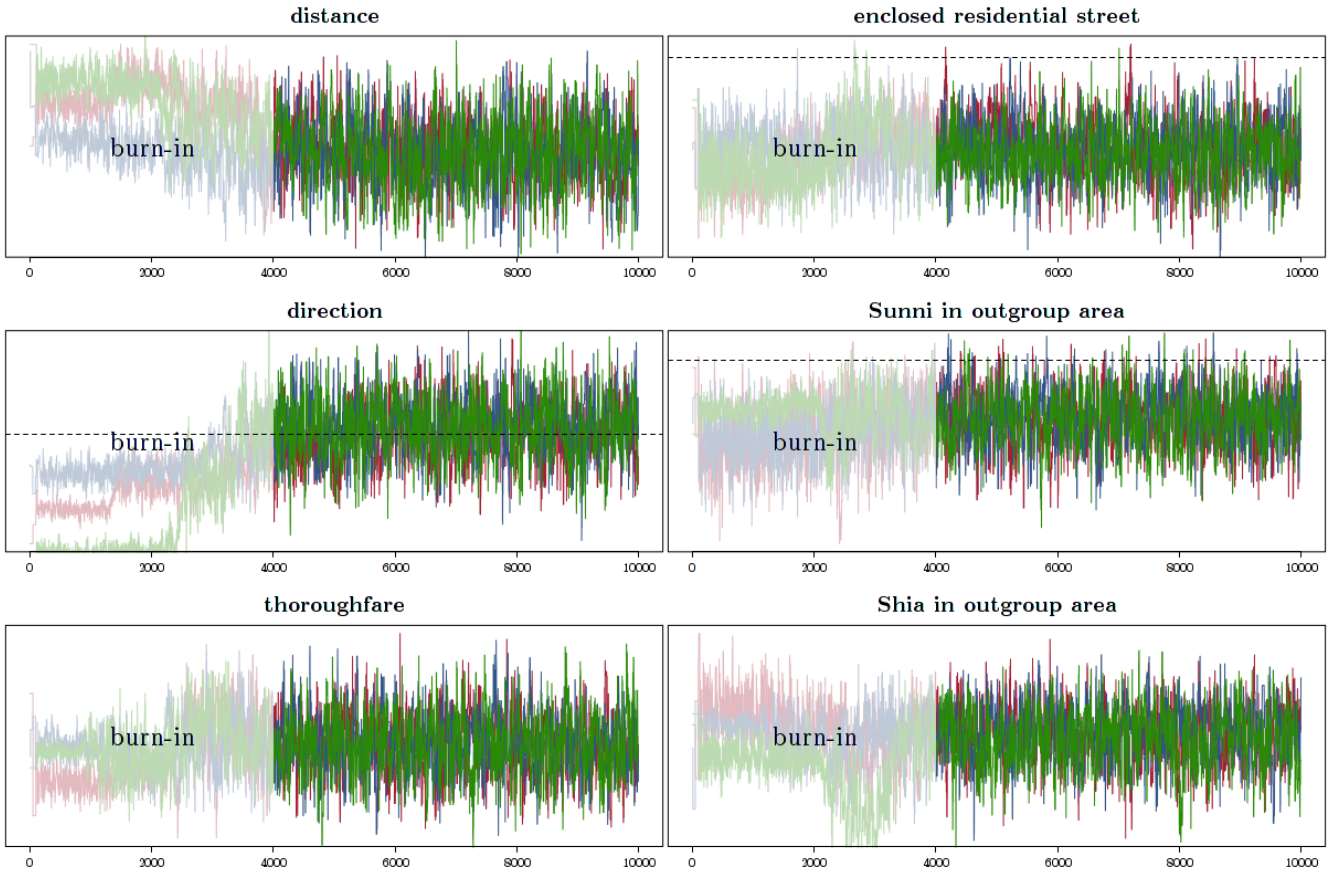

Figure 25: Traces of three chains, denoted by color. Visual inspection suggests that a burn-in of 4,000 iterations is adequate and that autocorrelation is low relative to chain length.

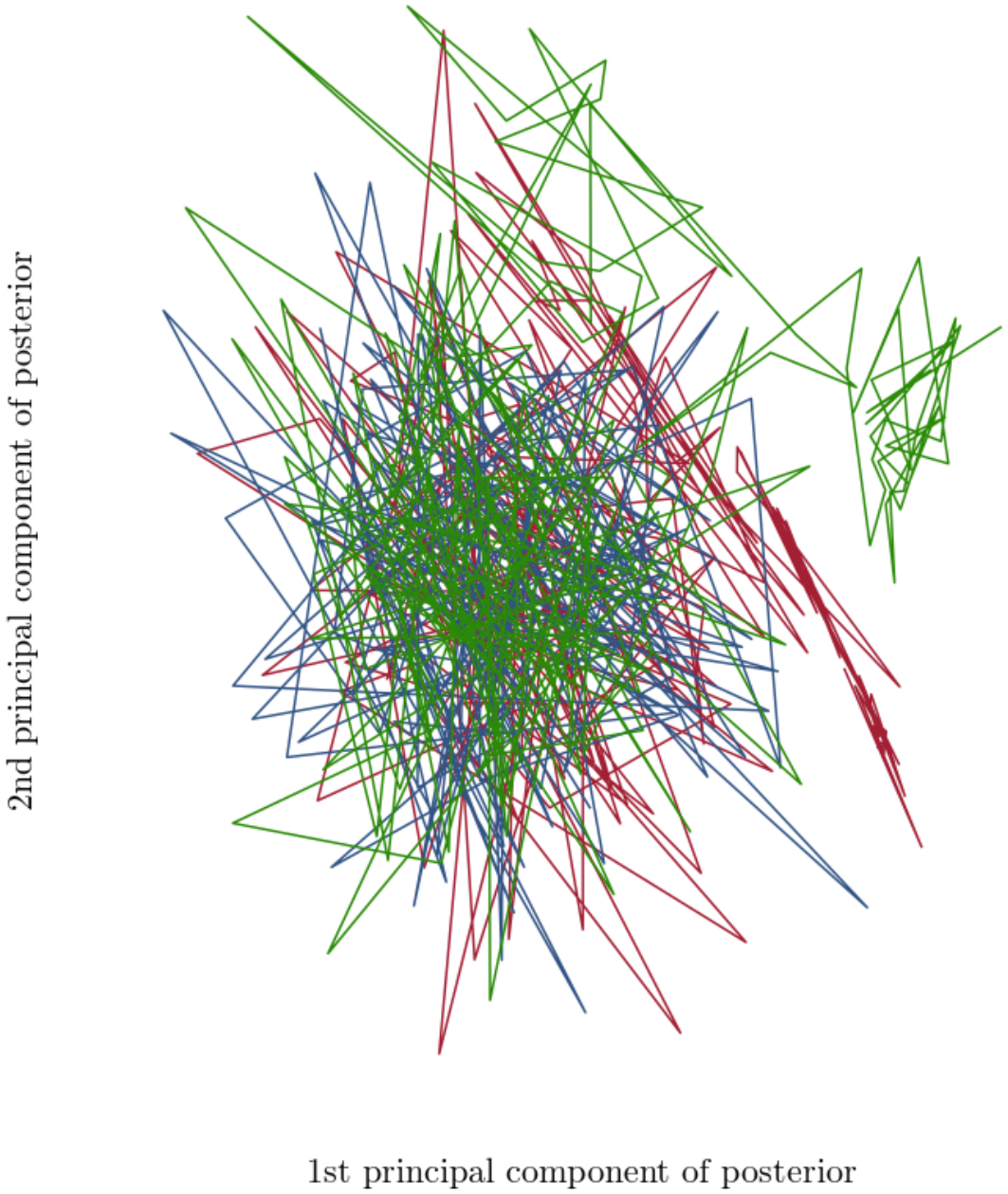

Figure 26: Visualizing MCMC chains with the first two principal components of the posterior distribution. The plot shows that chains converge to the same region in the parameter space, with excellent mixing.

References

Aronow, Peter M. and Cyrus Samii. n.d. "Estimating Average Causal Effects Under Interference Between Units." http://arxiv.org/pdf/1305.6156v1.pdf.

Aschauer, David Alan. 1989. "Is Public Expenditure Productive?" *Journal of Monetary Economics* 23:177–200.

Baker, III, James A., Lee H. Hamilton, Lawrence S. Eagleburger, Vernon E. Jordan, Jr., Edwin Meese, III, Sandra Day O'Connor, Leon E. Panetta, William J. Perry, Charles S. Robb and Alan K. Simpson. 2006. The Iraq Study Group Report. report U.S. Institute of Peace.

Banerjee, Abhijit, Esther Duflo and Nancy Qian. 2012. On the Road: Access to Transportation Infrastructure and Economic Growth in China. Technical Report 17897. http://www.nber.org/papers/w17897.

Baum-Snow, Nathaniel. 2007. "Did Highways Cause Suburbanization?" *Quarterly Journal of Economics* 122:778–805.

Baum-Snow, Nathaniel, Loren Brandt, J. Vernon Henderson, Matthew A. Turner and Qinghua Zhang. n.d. "Roads, Railroads and Decentralization of Chinese Cities.".

Bhat, Chandra R. 2001. "Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model." *Transportation Research Part B: Methodological* 35(7):677–693.

Blair, Graeme. 2016. On the Geography of Assets and Citizens: How Proximity to Oil Production Shapes Political Order PhD thesis Princeton University.

Bollobás, Béla. 1998. *Modern Graph Theory*. New York: Springer.

Bowers, Jake, Mark Fredrickson and Costas Panagopoulos. 2013. "Reasoning about interference between units: a general framework." *Political Analysis* 21:97–124.

Breslow, Norman E. 1974. "Covariance Analysis of Censored Survival Data." *Biometrics* 30(1):89–99.

Briggs, Ryan C. 2012. "Electrifying the base? Aid and incumbent advantage in Ghana." *The Journal of Modern African Studies* 50:603–624.

Burgess, Robin, Remi Jedwab, Edward Miguel, Ameet Morjaria and Gerard Padr i Miquel. 2015. "The Value of Democracy: Evidence from Road Building in Kenya." *American Economic Review* 105:1817–1851.

Carmody, Padraig. 2009. "Cruciform sovereignty, matrix governance and the scramble for Africa's oil: Insights from Chad and Sudan." *Political Geography* 28:353–361.

Casaburi, Lorenzo, Rachel Glennerster and Tavneet Suri. 2013. Rural Roads and Intermediated Trade: Regression Discontinuity Evidence from Sierra Leone. Technical report.

Caughey, Devin. 2012. Congress, Public Opinion, and Representation in the One-Party South, 1930s–1960s PhD thesis University of California, Berkeley.

Chandra, Amitabh and Eric Thompson. 2000. "Does Public Infrastructure Affect Economic Activity? Evidence from the Rural Interstate Highway System." *Regional Science and Urban Economics* 30:457–490.

Christia, Fotini and Dean Knox. n.d. "Geographic Segregation in Baghdad: Detecting Out-group Aversion in Walking Routes.".

Christia, Fotini, Dean Knox and Jaffar Al-Rikabi. n.d. "Networks of Sectarianism: Experimental Evidence on Access to Services in Baghdad.".

Christia, Fotini, Elizabeth Dekeyser and Dean Knox. n.d. "Gauging Shia Public Opinion: A Survey of Iranian and Iraqi Religious Pilgrims.".

Cohen, Jeffrey P. and Catherine J. Morrison Paul. 2004. "Public Infrastructure Investment, Interstate Spatial Spillovers, and Manufacturing Costs." *Review of Economics and Statistics* 86:551–560.

Converse, Philip E. 1962. "Information Flow and the Stability of Partisan Attitudes." *Public Opinion Quarterly* 26(4):578–599.

Cornfield, J. 1978. "Randomization by group: a formal analysis." *American Journal of Epidemiology* 108(2):100–102.

Cranmer, Skyler J. and Bruce A. Desmarais. 2011. "Inferential Network Analysis with Exponential Random Graph Models." *Political Analysis* 19(1):66–86.

Cranmer, Skyler J., Philip Leifeld, Scott D. McClurg and Meridith Rolfe. 2016. "Navigating the Range of Statistical Tools for Inferential Network Analysis." *American Journal of Political Science* Forthcoming:416–424.

Damluji, Mona. 2010. "'Securing Democracy in Iraq': Sectarian Politics and Segregation in Baghdad, 2003–2007." *Traditional Dwellings and Settlements Review* 21(2):71–87.

Del Bo, Chiara F. and Massimo Florio. 2012. "Infrastructure and Growth in a Spatial Framework: Evidence from the EU regions." *European Planning Studies* 20:1393–1414.

Dell, Melissa. 2015. "Trafficking Networks and the Mexican Drug War." *American Economic Review* 105(6):1738–1779.

Dodds, Peter Sheridan, Roby Muhamad and Duncan J. Watts. 2003. "An Experimental Study of Search in Global Social Networks." *Science* 301:827–829.

Donaldson, Dave. forthcoming. "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure." *American Economic Review* .

Eckstein, Harry H. 1975. Case studies and theory in political science. In *Handbook of Political Science*, ed. Fred J. Greenstein and Nelson W. Polsby. Reading, MA: Addison-Wesley.

Efron, Bradley. 1977. "The Efficiency of Cox's Likelihood Function for Censored Data." *Journal of the American Statistical Association* 72(359):557–565.

Elkins, Zachary and Beth Simmons. 2005. "On Waves, Clusters and Diffusion: A Conceptual Framework." *American Academy of Political and Social Science* 598:33–51.

Eubank, Nicholas. 2016. "Social Networks and the Political Salience of Ethnicity.".

Fogel, Robert. 1962. "A Quantitative Approach to the Study of Railroads in American Economic Growth: A Report of Some Preliminary Findings." *Journal of Economic History* 22(2):163–197.

Fogel, Robert. 1964. *Railroads and American Economic Growth: Essays in Econometric History*. Baltimore: Johns Hopkins Press.

Fosgerau, Mogens, Emma Frejinger and Anders Karlstrom. 2013. "A link based network route choice model with unrestricted choice set." *Transportation Research Part B: Methodological* 56:70–80.

Gerring, John. 2007. Is There a (Viable) Crucial-Case Method? Vol. 40 pp. 231–253.

Gremlin. 1976. *Blockade*. San Diego: Gremlin Industries. Arcade game.

Habyarimana, James, Macartan Humphreys, Daniel N. Posner and Jeremy M. Weinstein. 2007. "Why does ethnic diversity undermine public goods provision?" *American Political Science Review* 101(4):709–725.

Haines, Michael R. n.d. Historical, Demographic, Economic, and Social Data: The United States, 1790–2002 (dataset). Technical Report 2896. http://www.icpsr.umich.edu/icpsrweb/RCMD/studies/2896.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1):1–30.

Hartman, Alexandra C. and Benjamin S. Morse. n.d. "Wartime Violence, Empathy, and Altruism: Evidence from the Ivoirian Refugee Crisis in Liberia.".

Huckfeldt, Robert and John Sprague. 1987. "Networks in Context: The Social Flow of Political Information." *American Political Science Review* 81(4):1197–1216.

Huckfeldt, Robert and John Sprague. 1995. *Citizens, politics and social communication: Information and influence in an election campaign*. Cambridge: Cambridge University Press.

International Medical Corps. 2007. Iraqis on the Move: Sectarian Displacement in Baghdad. Technical report.

Iraq Body Count. 2016. "Documented Civilian Deaths from Violence." Database.

Killworth, Peter D. and H. Russell Bernard. 1978. "The Reversal Small-World Experiment." *Social Networks* 1:159–192.

King, Gary. 1989. "Event Count Models for International Relations: Generalizations and Applications." *International Studies Quarterly* 33(2):123–147.

Kranton, Rachel E. and Deborah F. Minehart. 2001. "A Theory of Buyer-Seller Networks." *American Economic Review* 91(3):485–508.

Lawler, Gregory F. 1999. Loop-Erased Random Walk. In *Perplexing Problems in Probability*, ed. Maury Bramson and Rick Durrett. Boston: Birkhäuser.

Lee, Frances E. 2000. "Senate Representation and Coalition Building in Distributive Politics." *American Political Science Review* 94:59–72.

Lee, Lung-Fei. 1992. "On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models." *Econometric Theory* 8:518–552.

Marin, Jean-Michel, Pierre Pudlo, ChristianP. Robert and RobinJ. Ryder. 2012. "Approximate Bayesian computational methods." *Statistics and Computing* 22(6):1167–1180.

Michaels, Guy. 2008. "The effect of trade on the demand for skill—Evidence from the Interstate Highway System." *Review of Economics and Statistics* 90:683–701.

Miguel, Edward and Mary Kay Gugerty. 2005. "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya." *Journal of Public Economics* 89(11–12):2325–2368.

Milgram, Stanley. 1967. "The Small-World Problem." *Psychology Today* 1(1):61–67.

Mironova, Vera and Sam Whitt. 2014. "Ethnicity and Altruism After Violence: The Contact Hypothesis in Kosovo." *Journal of Experimental Political Science* 1:170–180.

Morrison, Andrew R. 1993. "Violence or Economics: What Drives Internal Migration in Guatemala?" *Economic Development and Cultural Change,* 41(4):817–831.

Nall, Clayton. 2013. "The Road to Division: How Interstate Highways Caused Geographic Polarization.".

Nall, Clayton. 2015. "The Political Consequences of Spatial Policies: How Interstate Highways Facilitated Geographic Polarization." *Journal of Politics* 77:394–406.

National Interregional Highway Committee, U.S. 1944. Interregional Highways: Message from the President of the United States, transmitting a report of the National Interregional Highway Committee, outlining and recommending a national system of interregional highways. Technical report U.S. Government Printing Office.

OpenStreetMap. 2016. "OpenStreetMap data.".

Pearl, Judea. 2000. *Causality.* New York: Cambridge University Press.

Rephann, Terance and Andrew Isserman. 1994. "New Highways as Economic Development Tools: An Evaluation Using Quasi-Experimental Matching Methods." *Regional Science and Urban Economics* 24:723–751.

Roberts, Ben and Dirk P. Kroese. 2007. "Estimating the Number of s–t Paths in a Graph." *Journal of Graph Algorithms and Applications* 11(1):195–214.

Rogers, Todd and Masa Aida. 2013. "Vote Self-Prediction Hardly Predicts Who Will Vote, and Is (Misleadingly) Unbiased." *American Politics Research* 42(3):503–528.

Rosenbaum, Paul R. 2002. *Observational Studies.* New York: Springer-Verlag.

Rosenfeld, Bryn, Kosuke Imai and Jacob Shapiro. 2016. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." *American Journal of Political Science* 60(3):783–802.

Rubin, Donald B. 1991. "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism." *Biometrics* 47(4):1213–1234.

Samii, Cyrus. 2013. "Perils or Promise of Ethnic Integration? Evidence from a Hard Case in Burundi." *American Political Science Review* 107(3):558–573.

Schelling, Thomas C. 1969. "Models of Segregation." *American Economic Review* 59(2):488–493.

Schelling, Thomas C. 1971. "Dynamic Models of Segregation." *Journal of Mathematical Sociology* 1(2):143–186.

Tiebout, Charles M. 1956. "A Pure Theory of Local Expenditures." *Journal of Political Economy* 64(5):416–424.

Tripp, Charles. 2000. *A History of Iraq.* Cambridge: Cambridge University Press.

U.N. High Commissioner for Refugees. 2016. Stabilizing the situation of refugees and migrants in Europe. Proposals to the meeting of eu heads of state or government and turkey United Nations.

U.S. Bureau of Public Roads. 1955. General location of national system of Interstate highways, including all additional routes at urban areas designated in September 1955. Technical report U.S. Government Printing Office.

Valiant, Leslie G. 1979. "The Complexity of Enumeration and Reliability Problems." *Siam Journal of Computing* 8(3):410–421.

van der Laan, Mark J., Sandrine Dudoit and Sunduz Keles. 2004. "Asymptotic Optimality of Likelihood-Based Cross-Validation." *Statistical Applications in Genetics and Molecular Biology* 3(1):1–23.

Voigtlaender, Nico and Hans-Joachim Voth. 2014. Highway to Hitler. Technical Report 20150. http://www.nber.org/papers/w20150.

Wilson, David B. 1996. Generating Random Spanning Trees More Quickly than the Cover Time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing*. Association for Computing Machinery.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Young, H. Peyton. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton: Princeton University Press.

Zaller, John. 1989. "Bringing Converse Back In: Modeling Information Flow in Political Campaigns." *Political Analysis* 1(1):181–234.

Zhang, Junfu. 2004*a*. "A Dynamic Model of Residential Segregation." *Journal of Mathematical Sociology* 28(3):147–170.

Zhang, Junfu. 2004*b*. "Residential Segregation in an All-Integrationist World." *Journal of Economic Behavior & Organization* 54:533–550.