A Model for Path Data

Dean Knox[*]

This draft: August 21, 2015

Abstract

As network analysis grows more common, social scientists are increasingly faced with path data. Paths arise when actors strategically navigate toward a goal in a social, physical, or policy space. Highway networks, or collections of paths that connect major cities and intervening counties, are a prominent example. Existing statistical tools essentially assume that counties construct highway segments independently. I propose a new model that explicitly captures pathwise-dependence between observations, develop an estimation procedure, and demonstrate its properties. The random-path model (RPM) allows researchers to explore factors that shape the trajectories of paths. I evaluate the method by simulation and an empirical application.

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

1 Introduction

Paths arise in a vast array of social science settings: voters navigate social networks in search of information, planners route roads to connect cities, and nations trace trajectories in a policy space. Each of these examples is, at its core, a story about decision-making in pursuit of a long-term goal, while facing a constrained set of intermediate steps that depend on past choices and future options.

Much research in political science and economics has sought to explain why paths emerge in a particular way or to assess their impacts. For example, a substantial literature characterizes the role of distributive politics in spending on the U.S. Interstate Highway System, a collection of paths. In turn, highways have been linked to partisan migration and geographic polarization between urban and suburban areas (Nall, 2015). Researchers have increasingly recognized that standard models fail to account for the dependence between units that arises in this context—that is, whether a county is connected to the highway system depends in large part on whether neighboring counties are also connected.[1] I argue that paths are a unique class of outcomes and should be modeled accordingly, much as binary outcomes are commonly modeled with, say, logistic regression.

In this paper, I propose and demonstrate the properties of a new statistical model that provides a principled way to estimate parameters from observed paths. 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 likely larger than that of deviating from the shortest route. A path may be thought of as a set of dependent dyadic binary observations, where an outcome of 1 indicates that the dyad is directly connected by one step on the path, and any two units on the path are indirectly connected by a series of such steps. A more complete definition is provided in section 3. In the random-path model (RPM), paths are assumed to be drawn from a random-walk distribu-

[1] In particular, whether a county is connected to an east-west highway depend on whether its neighbors to the east and west are connected as well as whether its neighbors to the north and south are *not* connected.

tion that has been *a priori* conditioned to prevent visiting any point twice. While random-walk distributions are well-studied, the additional no-loop constraint poses a challenge for estimation in that it makes the resulting probability mass function intractable. 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. Finally, I discuss extensions including RPMs for assignment of path-like treatments, thereby allowing causal inferences about their effects (Rubin, 1991).

In the remainder of this paper, I first briefly discuss several motivating examples 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 method by simulation and then presents a simple empirical application. Section 5 concludes.

2 Motivation

Paths arise naturally in a variety of social and physical contexts—for social scientists, perhaps most notably when when agents purposefully navigate a social or geographic environment. A probabilistic model of path formation is appropriate when agents do so with limited information or otherwise bounded rationality. As a concrete example, Habyarimana et al. (2007) showed that co-ethnic social networks increase “findability” of strangers as part of their study of public goods provision. In their experiment, randomly selected Ugandan “runners” were given photographs of targets and asked to locate them within 24 hours, which they did with startling levels of success. Based on qualitative interviews with runners and higher success rates for co-ethnic targets, Habyarimana et al. conclude that runners used an ethnicity-based graph-traversal strategy to determine the paths by which they searched the Kampala social structure. Their findings provoke a number of follow-on questions: How does ethnicity compare to, say, religion as runners contemplate their next step? Are these strategies efficient, or do they lead to longer search times and lower success rates? How deeply ingrained are ethnicity-based strategies—do runners adapt when they are ineffective? A fuller model of these strategies would have implications not only for theories of social sanctioning, but also political information-seeking behavior (Huckfeldt and Sprague, 1995) and buyer-seller network formation (Kranton and Minehart, 2001). However, existing tools are unable to capture the forward-looking behavior of runners and other agents, who choose their next step based on heuristics about how much closer it will bring them to a goal.

Related methodological issues arise in the distributive politics literature, where spending on transportation networks or electrical grids has received substantial attention as an outcome to be explained. Prior research has tended to ignore the spatial element of infrastructural policy, instead modeling it as a purely local pork-barrel or developmental spending (see, e.g. Burgess et al., 2013). Given that highway networks 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. In this context, analyses that ignore the spatial dependence of connective infrastructure are as much of an "exercise in self-deception" as those that ignore cluster randomization (Cornfield, 1978). However, existing models do not permit principled testing of hypotheses about factors shaping the trajectories of paths.

More recently, scholars have begun to explore the effects of path-assigned treatments with innovative research designs that partially address inferential challenges in these settings. By subsetting to suburban U.S. counties and assuming that the direction in which Interstate highways cross major cities is ignorable, Nall (2013, 2015) shows that highways facilitated Republican migration out of urban centers and induced "geographic polarization," with continuing implications for American political geography. Voigtlaender and Voth (2014) use a nationwide highway suitability analysis as an instrument for the German *Autobahn* network, demonstrating that construction spending and employment dramatically reduced opposition to the Nazi regime. In these applications, the authors deploy context-specific knowledge of the path-assignment mechanism to gain leverage on their counterfactual questions. These designs seek to approximate an ideal experiment in which a number of candidate highways are proposed, but only some are randomly selected for construction. RPM takes this intuition and generalizes it for cases in which such fine-grained data is unavailable. Thus, it is well-suited for incorporation into Bayesian multilevel models or causal inferences that involve path-like phenomena.

3 Model

In this section, I briefly discuss two "views" of random walks. 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. Finally, I contrast the proposed method with existing approaches to related problems, including spatial regression models and exponential random graph models.

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.[2] 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 of linear predictors. Assume edge weights can be written as

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

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

$$\mathcal{L}_{\text{walk}}(\beta \mid \boldsymbol{X}, \gamma) = \prod_{t=0}^{k-1} \frac{\exp\left([\boldsymbol{X}_m \beta^m]_{\gamma_t, \gamma_{t+1}}\right)}{\sum_{j'} \exp\left([\boldsymbol{X}_m \beta^m]_{\gamma_t, j'}\right)}$$

[2] 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 $E_{v_t, v_{t+1}} > 0$.

Recent developments in estimation of the random-walk model, primarily in the transportation literature, are discussed in section 3.4.

3.2 Random Path Distribution as Conditional Random Walk

The random walk, while analytically tractable, is an unlikely model for many social phenomena. In particular, random walkers are neither forward- or backward-looking, and under typical conditions they are likely to revisit many nodes. This maps poorly to, e.g., highways, which are designed by planners who seek to minimize some cost function under the constraints of bounded rationality.

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 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 corresponding random-path likelihood is simply

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

This random-path distribution exhibits oracle properties that make it well-suited for modeling the strategic behavior discussed in section 2. 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" and more "cheaply."

3.3 Estimation

I begin by presenting an algorithm for evaluating the RPM likelihood exactly. 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 RPM Likelihood

The chief difficulty in evaluating equation 1 is its denominator, which varies with β and involves summing over the set of all possible paths between the observed start- and endpoints, γ_0 and γ_k. The maximum likelihood estimate of β, for example, involves maximizing the unconditional (random-walk) probability of the observed path relative to the totaled random-walk probabilities of every other path that could have been drawn. In algorithm 1, I present an exact method that explicitly enumerates every possible path by depth-first search, then sums the random-walk probabilities of mutually exclusive paths.

In practice, explicit enumeration of all possible paths between two nodes is computationally infeasible for moderately sized or dense graphs. For example, in complete graphs, where every node is connected to every other, the number of possible paths between any two nodes is given by $K(N) = \sum_{k=0}^{N-2} \frac{(N-2)!}{k!}$. Even in a ten-node complete graph, 109,601 paths are possible. Building on the intuition behind the exact approach, I develop algorithm 2 to approximate the likelihood function to arbitrary precision.

A common numerical approach to summations over hard-to-enumerate domains is Monte Carlo

```
Data:
  starting node $\gamma_0$, terminus $\gamma_k$, covariates $\boldsymbol{X}$, parameters $\beta$

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

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

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

Algorithm 1: 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.

integration. For example, the integral $\int_a^b e^{-\alpha x^2}$ can be approximated by randomly sampling points on the uniform $[a, b]$ distribution, evaluating the integrand at each point, averaging the results, and multiplying by the size of the sampling space $(b - a)$. Let Ψ be the uniform distribution over the set of all possible paths $\mathcal{P}$. The directly analogous approach to estimating the denominator of equation 1 is to repeatedly draw $\psi \sim \Psi$, evaluate the summand for each sampled element, average to estimate $\mathbb{E}_\Psi[\Pr(\Gamma = \Psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)]$, and multiply the average by the total number of paths $|\mathcal{P}|$. Because the numerator is calculated exactly, the simulated likelihood (Lee, 1992) inherits the desirable property of converging to the exact likelihood in algorithm 1 (up to a multiplicative constant) as the number of samples (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 the uniform distribution draws each path converges to $\frac{S}{|\mathcal{P}|}$ as S grows large.

There are two complications in this procedure. First, counting the total number of paths is an #P-hard problem (Valiant, 1979), and the fastest known algorithm for calculating it is depth-first search.[3] However, $|\mathcal{P}|$ does not involve the RPM parameters, β, and can thus be absorbed into the normalizing constant of the RPM likelihood function. The RPM likelihood is then given by

$$\mathcal{L}_{\text{path}}(\beta \mid \boldsymbol{X}, \gamma) \;\; \propto \;\; \frac{\prod_{t=0}^{k-1} \frac{\exp\left([\boldsymbol{X}_m \beta^m]_{\gamma_t, \gamma_{t+1}}\right)}{\sum_{j'} \exp\left([\boldsymbol{X}_m \beta^m]_{\gamma_t, j'}\right)}}{\mathbb{E}_\Psi[\Pr(\Gamma = \Psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)]}. \tag{2}$$

Equation 2 could be approximated by the simulated likelihood method, through Monte-Carlo sampling from Ψ. Unfortunately, there is no known way to sample from Ψ directly. To deal with this, I adapt a non-uniform distribution for importance sampling on $\mathcal{P}$. Two options for sampling non-uniformly from $\mathcal{P}$ are the self-avoiding walk (SAW) and the loop-erased random walk (LERW). The SAW is a 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.[4] The LERW begins with a pure random walk, then retraces its steps and removes loops that return to previously visited nodes. (The procedure is defined formally as part of algorithm 2.) 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

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

[4] 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 target 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.

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

$$\begin{aligned}&\hat{\mathbb{E}}_\Psi[\Pr(\Gamma = \Psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)] \\ &\quad = \sum_{s=1}^{S} \Pr(\Gamma = L_s \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)\ w(L_s),\end{aligned} \tag{3}$$

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 likelihood 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.1. 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 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; the full procedure is given in algorithm 2.

```
Data:
  starting node $\gamma_0$, teminus $\gamma_k$, covariates $\boldsymbol{X}$, parameters $\beta$
  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
  $\hat{\mathbb{E}}_\Psi[\Pr(\Gamma = \Psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)]$,
    expected random-walk probability of uniformly sampled path

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 $\hat{\mathbb{E}}_\Psi[\Pr(\Gamma = \Psi \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta)] =$
        $\frac{1}{\sum_{l=1}^{s} w_s} \sum_{l=1}^{s} w_s \Pr\left(\Gamma = \psi_s \mid v_0 = \gamma_0, v_K = \gamma_k, \boldsymbol{X}, \beta\right)$
    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 $\psi = (\gamma_0)$,   $i = \gamma_0$
    while $i \neq \gamma_k$ do
        sample $j$ uniformly from $\mathcal{N}_i$
        step to $i = j$ and append to $\psi$
    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 $\psi$
        end
        $t$+=1
    end
    return $\psi$
```

Algorithm 2: 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.

3.3.2 Estimation by MCMC

The simulated likelihood function in equation 2 forms the basis for statistical inference. Neither it nor the true likelihood function to which it converges are necessarily well-behaved. For example, it is possible to construct examples with multi-peaked likelihoods, although they have not been found in the applications explored here.

For the Hawai'i simulation in section 4, I estimate RPM by Metropolis–Hastings (MH). Such Markov chain Monte Carlo (MCMC) methods repeatedly evaluate the simulated likelihood function, and each evaluation is computationally expensive. I discuss some obvious steps toward program optimization that greatly reduce running time. The full MCMC procedure is formally outlined in algorithm 3 and informally discussed below.

First, the sampling-reweighting procedure involves large numbers of simulated paths and matrix determinants. Rather than repeating algorithm 2 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 acceptance ratio, as the simulated likelihood of both current and proposed parameters are estimated with the same path-set.

Second, to evaluate the likelihood at any point in the parameter space, algorithm 2 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 $\gamma_0$, teminus $\gamma_k$, covariates $\boldsymbol{X}$
  unweighted graph $\tilde{G}$, number of path simulations $S$
  initial parameters $\beta^{(0)}$, gridded parameter space $\tilde{\boldsymbol{\mathcal{B}}}$
  number of Metropolis-Hastings samples $R$, proposal distribution $Q(\beta^*; \beta^{(t)})$

Result:
  $R$ correlated samples from posterior of parameters $\beta$

Algorithm ChainMH($\gamma$, $\boldsymbol{X}$, $\beta^{(0)}$, $\tilde{\boldsymbol{\mathcal{B}}}$, $Q$)
    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 $\beta^* \sim Q(\beta^*; \beta^{(r)})$
        calculate acceptance ratio $\alpha = \frac{\texttt{ApproxSimLikelihood}(\beta^*)}{\texttt{ApproxSimLikelihood}(\beta^{(r)})}$
        if $\alpha < 1$ and jump $\sim$ 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($\beta$)
    set $\tilde{\beta}$ to center of grid cell in $\tilde{B}$ containing $\beta$
    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 $\hat{\mathcal{L}}(\tilde{\beta} \mid \boldsymbol{X}, \gamma) = \frac{\Pr\left(\Gamma=\gamma \mid v_0=\gamma_0, v_K=\gamma_k, \boldsymbol{X}, \tilde{\beta}\right)}{\sum_{l=1}^{s} w_s \Pr\left(\Gamma=\psi_s \mid v_0=\gamma_0, v_K=\gamma_k, \boldsymbol{X}, \tilde{\beta}\right)}$
    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.

3.4 Comments

Here, I discuss RPM in the context of related methods. The methodological challenges that arise in RPM are closely related to those in conditional maximum likelihood estimation, e.g. for the fixed-effects logistic regression model, where conditioning on a sufficient statistic to circumvent the incidental parameters issue induces a combinatorics problem in the denominator of the conditional likelihood. In contrast to the fixed-effect logit setting, the combinatorics in the denominator are sufficiently complex that no convenient approximations (such as the Efron or Breslow methods) are currently known.

A similar difficulty arises in population genetics, where analytical formulae for the likelihood are unavailable or computationally expensive to evaluate, but simulations from the model are cheap. Rubin (1984) considered an accept-reject method for "likelihood-free" MCMC, in which a parameter proposal is made, a sample is simulated with those parameters, and the proposal is accepted if the simulated sample "matches" the observed sample on some sufficient statistic. Considerable progress has been made in recent years on approximate Bayesian computational (ABC) methods as increasingly complex models become feasible (see Marin et al., 2012, for a survey of recent developments). The method employed in this paper, which uses Monte Carlo simulations to integrate over the outcome space, is more closely related to ABC methods than typical simulated likelihood applications, which use Monte Carlo integration to marginalize nuisance variables such as random coefficients (Bhat, 2001). However, use of a simulated likelihood allows the RPM estimation procedure to sidestep common problems in ABC, such as when only "quasi-sufficient" statistics are available, or when the matching space is continuous so that a "tolerance" is required to address the zero probability of exact matches.

The random-walk model presented in part 3.1 is considered in the transportation literature, where route choices[5] are often modeled by recursive multinomial logit. Frejinger, Bierlaire and

[5] Somewhat confusingly referred to as "paths" in the transportation context, despite modeling choices that allow cycles; these correspond to "walks" in the graph context.

Ben-Akiva (2009) show that traditional estimation procedures are biased and derive a corrective walk-sampling procedure for the random-walk model that is similar in motivation to the LERW reweighting scheme of algorithm 2. Fosgerau, Frejinger and Karlstrom (2013) consider the problem that these models allow loops, including infinite loops; they report that such events are rare in their particular graph structure. Mai, Frejinger and Fosgerau (n.d.) propose an extension that incorporates forward-looking behavior into the random-walk model.

RPMs, like exponential random graph models (ERGMs), are models of graph formation. They differ in that ERGMs model dyadic link formation, such as international conflict, allowing for the influence of local network structure (in political science, see Cranmer and Desmarais, 2011). For example, war between two countries may increase the probability of involvement by their respective allies. In contrast, RPMs are influenced by network structure over longer ranges, are subject to more severe constraints, and generally address a different class of questions.

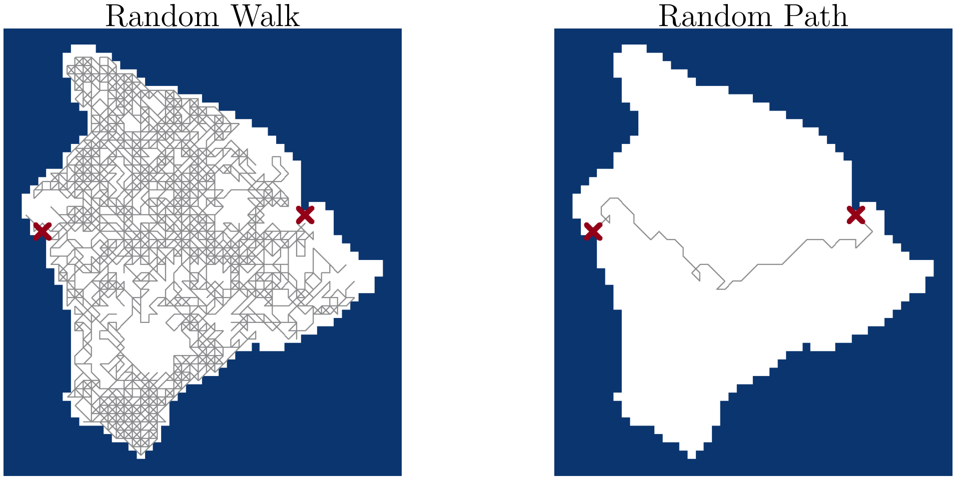
Lastly, a natural question is whether existing spatial or graph-formation models can produce results similar to RPMs. I briefly discuss options for spatial modeling, drawing heavily on Dormann et al. (2007), in the context of a geographic RPM. To fix ideas, consider a road that must pass through a city on the central cell of a 3×3 square grid, where the surrounding eight cells are suburbs—in essence, the setup in Nall (2013)—and the outcome is a binary road indicator. Assuming a preference for straight roads, the local autocorrelation structure will be positive for directly opposing cells (e.g., north–south, northeast–southwest) and negative for adjacent cells (e.g., east–southeast). This is difficult to capture with standard spatial models, which typically assume isotropy, or equal autocorrelation in all directions. While anisotropic models are also possible, these allow for elliptical autocorrelation (stronger in certain directions) at best, rather than the cross-shaped structure that arises naturally here. Spatial models generally also assume stationarity, or that the dependence structure persists even as the grid grows and distance from the urban center becomes larger. Allowing for changes in space—for example, tightening local dependence in a mountain pass—would be difficult, if not impossible, without ad-hoc model specifications. Generalized least squares is infeasible because the covariance matrix is clearly not positive definite, and spatial models that require partitioning nodes into clusters are inapplicable. More generally, the spatial models discussed here necessarily assume that dependence decays with distance at some estimated rate. The Hawai'i example in section 4, in which a road detours around a mountain so that cells on opposite sides have a strong negative correlation despite their distance, shows that this is a poor modeling assumption for path data. Finally, spatial autocovariate and autoregressive models essentially treat local dependence as a nuisance. As a result, they are well-suited to describing the association between non-spatial covariates and an outcome, but unable to estimate interpretable interactions or simulate counterfactual roads, as a model of treatment assignment requires.

4 Applications

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 iterated to assess the consistency of the estimation procedure. Finally, I describe a small empirical application to U.S. Interstate highways in Kentucky.

4.1 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 1 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.



Random Walk

Random Path

Figure 1: 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."[6] The covariates are shown in figure 2. For a walker at cell i, the unconditional (random-walk) probability of stepping to adjacent cell j is

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

The random-path distribution is the conditional random-walk distribution, given that the

[6] 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.

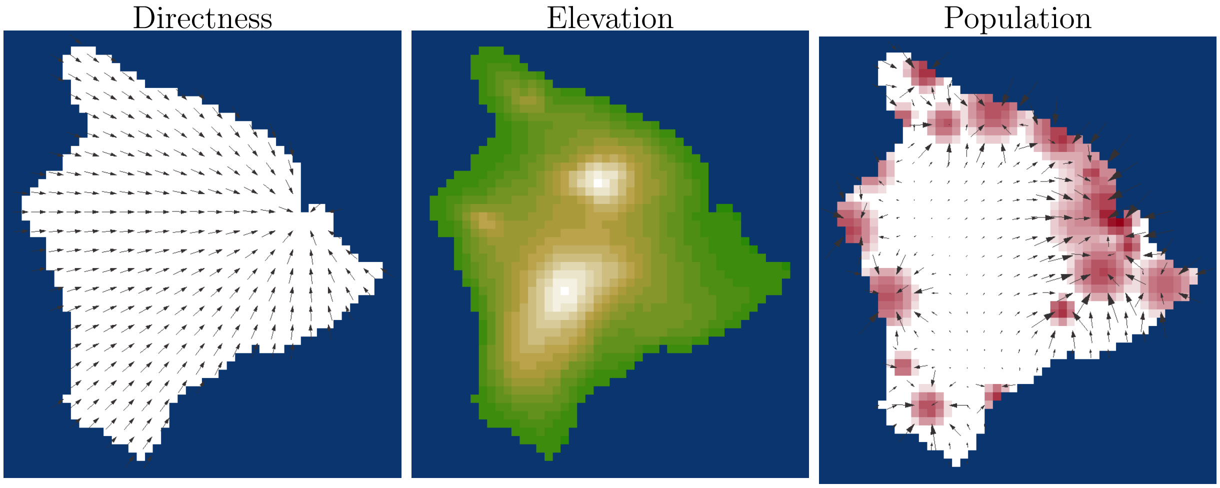


Figure 2: RPM covariates on a 50×50 grid. Direction toward target (left) indicated by arrows. Elevation (center) in terrain colors, green at sea level and white at $\sim 4,000$ m, around the peaks of Mauna Kea and Mauna Loa. Log-population (right) plotted in red, with higher density in more opaque regions. Arrows show the direction of population gravitational pull, with arrow size indicating force.

walk does not contain cycles. The simulation distribution of a random-path model is the result 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 3, I show the result of increasing β_{dir}. The left panel in figure 3 is a larger sample from the baseline distribution with all parameters set to zero (the same RPM that generated the right panel of figure 1). 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 4 and 5 depict the effects of β_{elev} and β_{pop}, respectively.

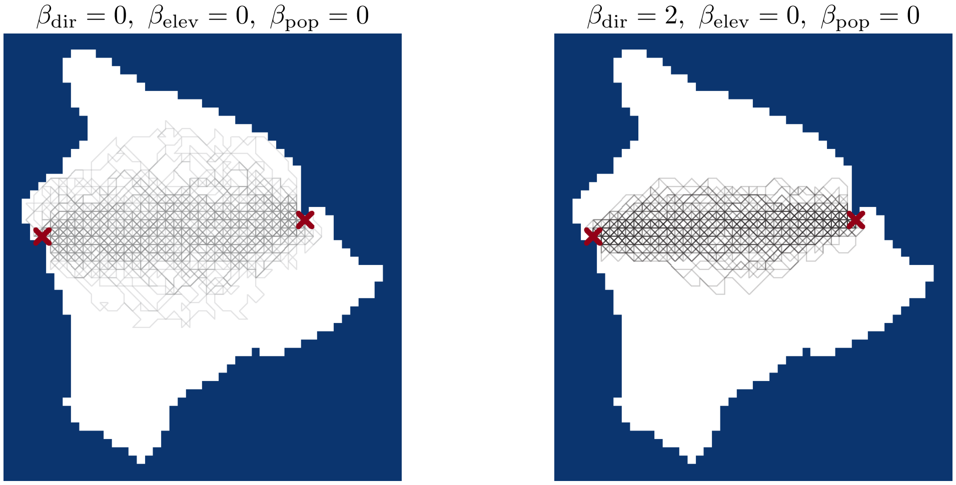
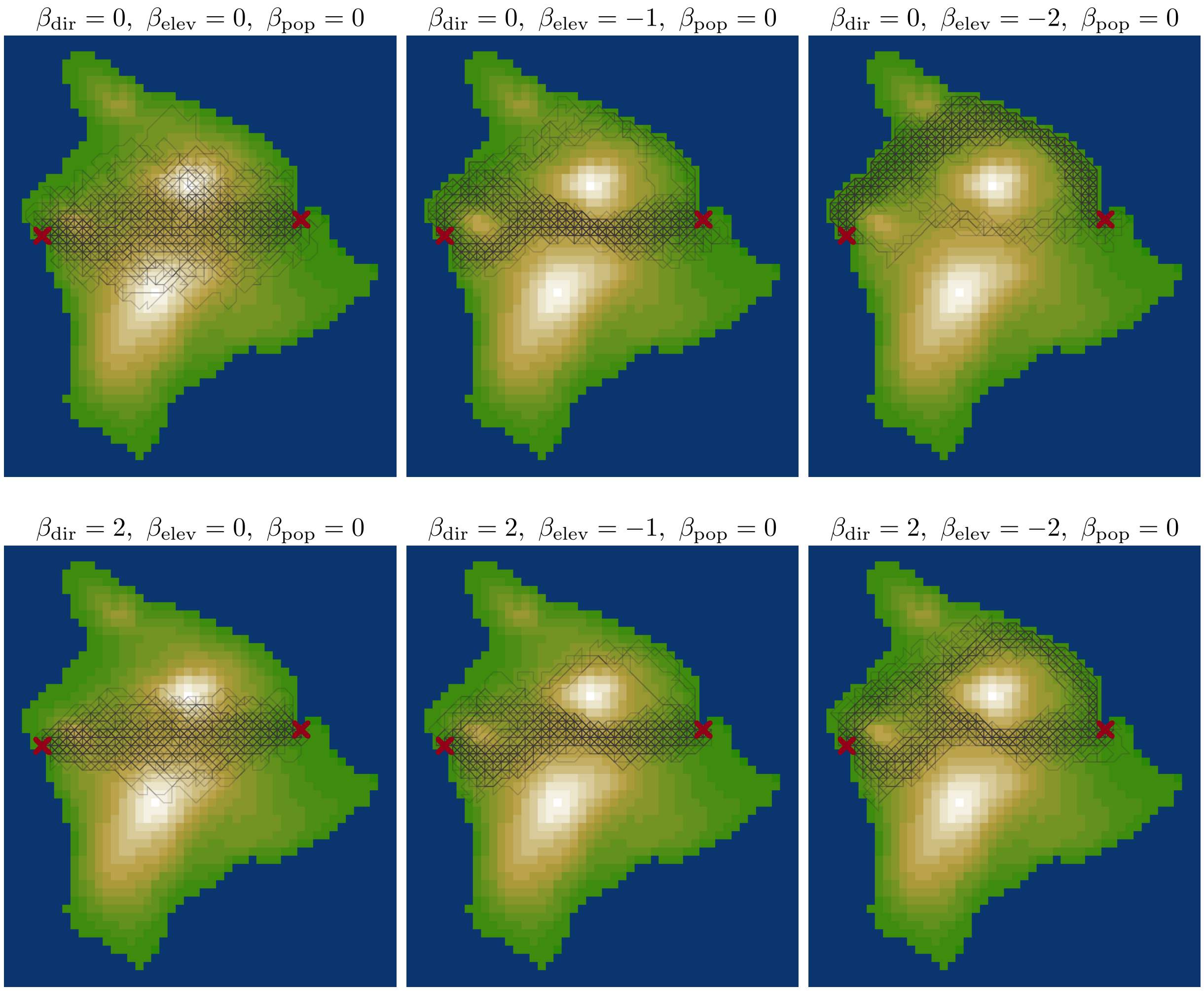

$\beta_{\text{dir}} = 0,\ \beta_{\text{elev}} = 0,\ \beta_{\text{pop}} = 0$
$\beta_{\text{dir}} = 2,\ \beta_{\text{elev}} = 0,\ \beta_{\text{pop}} = 0$

Figure 3: Higher values of β_{dir} (right) result in a tighter distribution with more direct paths than the baseline (left).



$\beta_{\text{dir}} = 0,\ \beta_{\text{elev}} = 0,\ \beta_{\text{pop}} = 0$

$\beta_{\text{dir}} = 0,\ \beta_{\text{elev}} = -1,\ \beta_{\text{pop}} = 0$

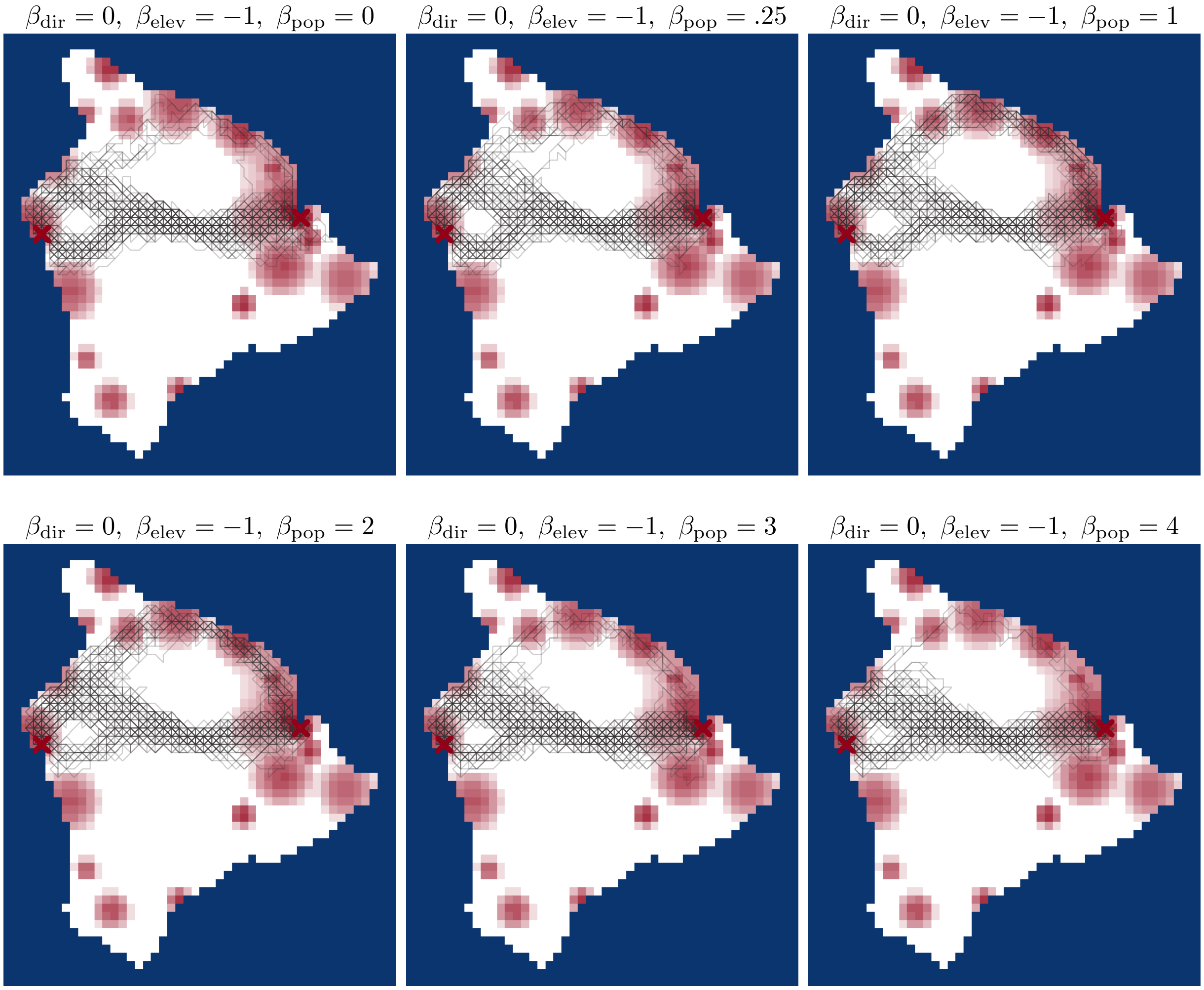
$\beta_{\text{dir}} = 0,\ \beta_{\text{elev}} = -2,\ \beta_{\text{pop}} = 0$

$\beta_{\text{dir}} = 2,\ \beta_{\text{elev}} = 0,\ \beta_{\text{pop}} = 0$

$\beta_{\text{dir}} = 2,\ \beta_{\text{elev}} = -1,\ \beta_{\text{pop}} = 0$

$\beta_{\text{dir}} = 2,\ \beta_{\text{elev}} = -2,\ \beta_{\text{pop}} = 0$

Figure 4: 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.



$\beta_{\text{dir}} = 0,\ \beta_{\text{elev}} = -1,\ \beta_{\text{pop}} = 0$
$\beta_{\text{dir}} = 0,\ \beta_{\text{elev}} = -1,\ \beta_{\text{pop}} = .25$
$\beta_{\text{dir}} = 0,\ \beta_{\text{elev}} = -1,\ \beta_{\text{pop}} = 1$
$\beta_{\text{dir}} = 0,\ \beta_{\text{elev}} = -1,\ \beta_{\text{pop}} = 2$
$\beta_{\text{dir}} = 0,\ \beta_{\text{elev}} = -1,\ \beta_{\text{pop}} = 3$
$\beta_{\text{dir}} = 0,\ \beta_{\text{elev}} = -1,\ \beta_{\text{pop}} = 4$

Figure 5: 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).

4.2 Validating the Estimation Procedure

4.2.1 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 6) captures this 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 $[1, k]$.

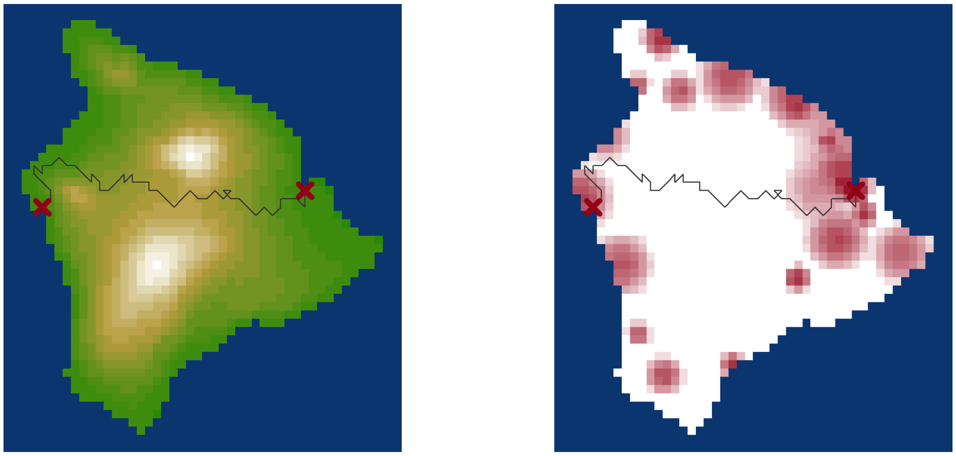


Figure 6: A single draw from RPM($\beta_{\text{dir}} = 0, \beta_{\text{elev}} = -1, \beta_{\text{pop}} = 0.5$), plotted against elevation (left) and population (right).

I evaluate the mixing of MH-sampled MCMC and the resulting estimates. Chain length was 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 7; elevation was estimated to be negative and correctly signed, while all other parameter estimates were insignificant.

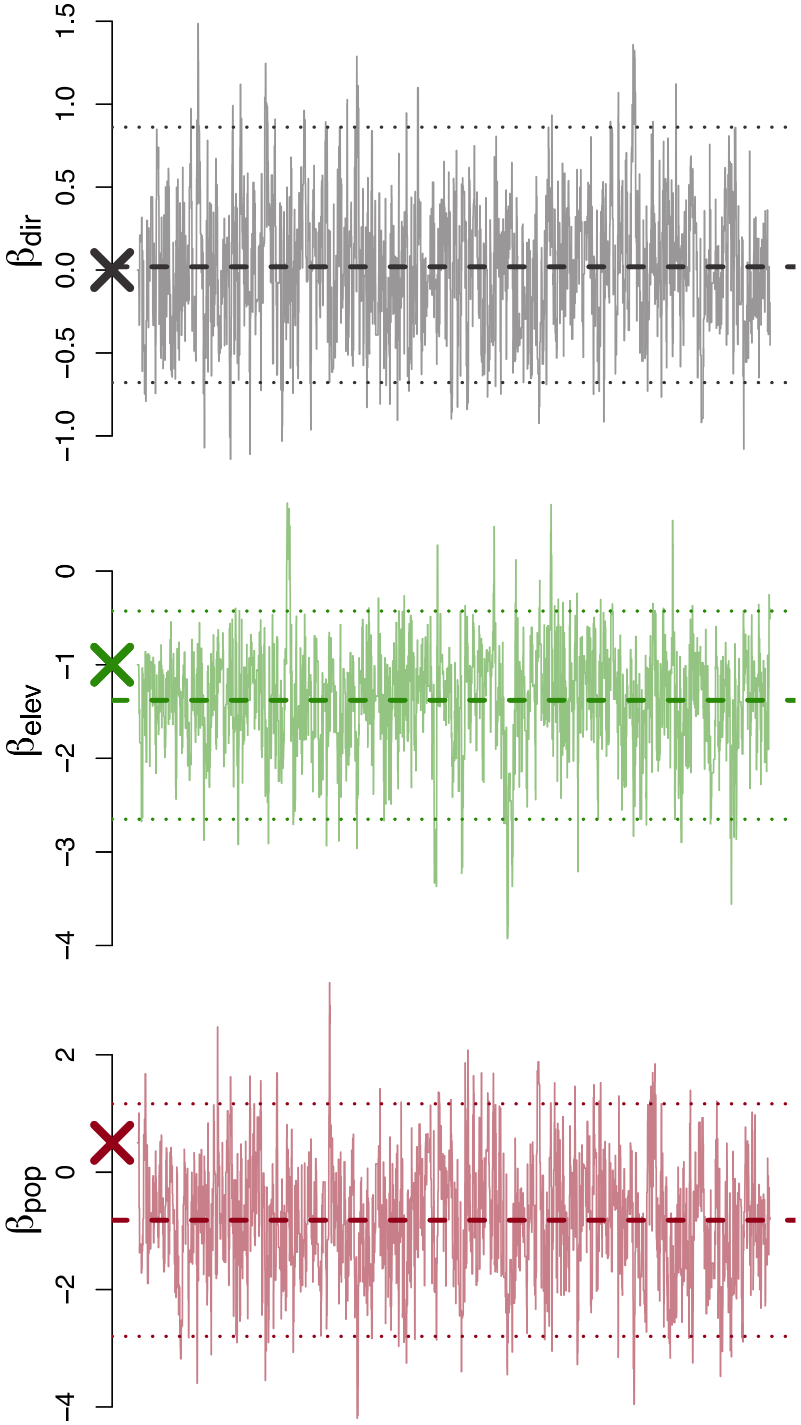


Figure 7: RPM posterior for the path depicted in figure 6, with sampled parameter values on the vertical axis and iterations on the horizontal. Parameter posterior means (dashed) and 95% marginal posterior credible intervals (dotted) are plotted horizontally over each chain. The true parameter is marked with a "×" on the vertical axis.

4.2.2 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 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 $\beta_{\text{elev}} = -2$.

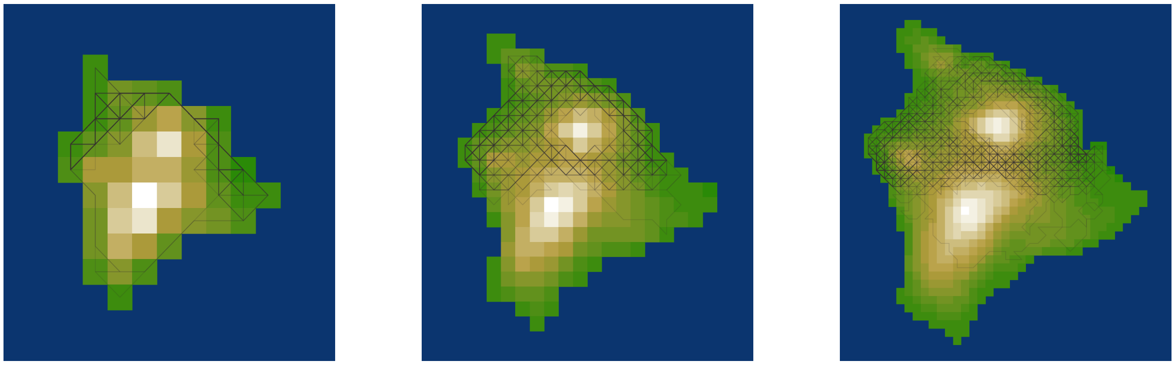


Figure 8: 100 draws from RPM($\beta_{\text{elev}} = -2$) on 10×10, 20×20, and 40×40 square grids

The true RPM distributions are shown in figure 8 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 9. 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.

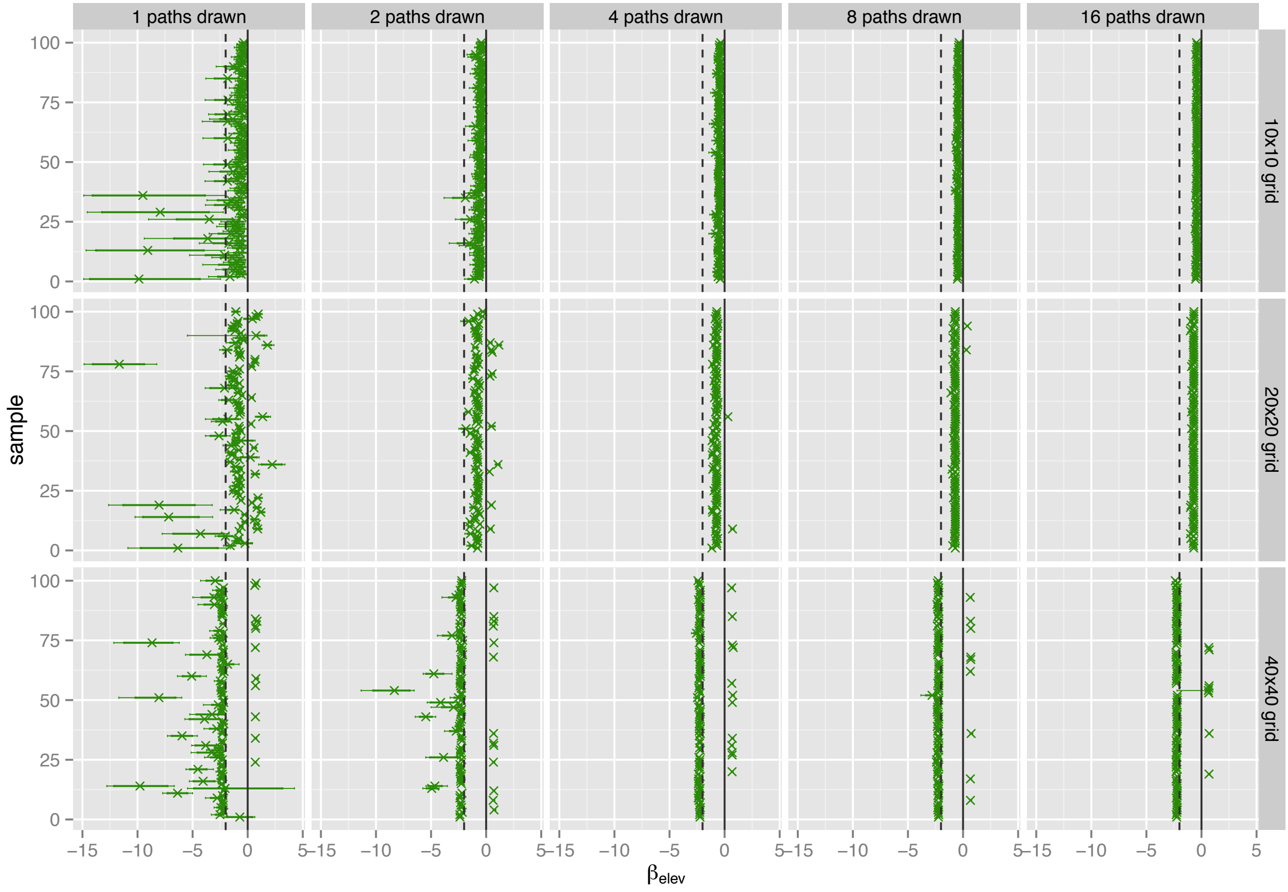


Figure 9: For each combination of sample size and grid size, 100 samples were drawn. Posterior means are marked with "×", 95% credible intervals with thin horizontal green lines, and 80% intervals with thick green horizontal lines. The true parameter is shown with a vertical black dotted line. Results 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.

4.3 Empirical Application

Next, I apply the RPM to a subset of the U.S. Interstate Highway System, often called the "greatest public works project in history." The Interstates, estimated to cost over 425 billion 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 (U.S. National Interregional Highway Committee, 1944) and explicitly stated decision criteria are publicly available. According to official documents, planners "base[d] its selection of routes primarily upon... interconnection of important [national and regional] cities," and secondarily, connecting smaller urban centers and counties with high agricultural product "as practicable." While RPM estimates are consistent with this account of highway construction, I show that distributive politics and representation play a substantial, statistically significant, and previously unmentioned role: for a Kentucky county, gaining hometown representation on the state Postwar Planning Committee would increase the probability of highway construction roughly as much as increasing county population by 600%.

It is reasonable to ask whether a stochastic model is appropriate for such a high-profile project, given the obvious benefits of deterministic optimization. However, algorithms for optimal spanning trees had only just been discovered when initial planning documents were published (U.S. Public Roads Administration, 1939), and there is no mention of their use in later reports even when other heuristics are thoroughly discussed. In addition, planning documents suggest a number of as-if-random sources of variation in trajectory, such as deviations for aesthetic reasons and periodic curves to increase the wakefulness of drivers.

Besides naming distance, population, and agricultural product as key design considerations, U.S. National Interregional Highway Committee (1944) also downplays the role of other explanatory variables. For example, topography was found to be important "in remarkably few places," and proposals for co-location with (or upgrading of) existing highways were rejected due to the difficulty and cost of widening existing right-of-ways to Interstate standards. Other, similar statements help further winnow the list of potential covariates, dramatically simplifying the RPM model.

The geographic unit of analysis was clearly defined by U.S. National Interregional Highway Committee to be cities and counties; I collapse cities into the county in which they are located. Counties are coded as receiving a highway (binary outcome) based on the National Transportation Atlas Database, after subsetting to initially planned Interstates as reported in U.S. Bureau of Public Roads (1955)—specifically, I-24, I-64, I-65, I-71, and I-75. Distances between counties are calculated based on the location of the county seat, and two counties are coded as "adjacent" if a straight line between their county seats does not pass within 10 miles of a third county seat.[7] The resulting graph is shown in figure 12. I assume that the intersections in Lexington and Louisville were politically predetermined, consistent with the goal of the "interconnection of important cities" and stated population thresholds. I further assume that border crossings into other states are fixed; while Kentucky legislators may have been able to affect crossing points, this influence is difficult to gauge without also modeling the decision-making of neighboring states. Fixing the crossings effectively eliminates variation that might otherwise be explained by the covariates, making parameter estimates noisier. This leaves seven highway segments with fixed endpoints. To address the undefined directionality of highways, I simulate the likelihood in both directions.

The economic covariates used are (1) distance, $\texttt{dist}_j$, as defined above; (2) 1940 log-population, $\texttt{pop}_j$, and (3) 1939 log-agricultural product, $\texttt{agp}_j$. The latter two were taken from (Haines, n.d.) and are the same variables and years explicitly discussed in U.S. National Interregional Highway Committee (1944). In addition, to assess the possibility of political influence on highway trajectories, I collected data on (4) the hometowns of the Kentucky Postwar Planning Committee members in 1945, $\texttt{plan}_j$, an influential state government body responsible for public works projects in the

[7] This is to allow highways to pass directly between diagonally touching county-pairs; it does not exclude any county-pairs that could reasonably be considered "adjacent." By recognizing additional edges, I avoid coding a county as "visited" by a highway segment when that segment would barely clip its corner. This does not preclude the county in question from being directly visited.

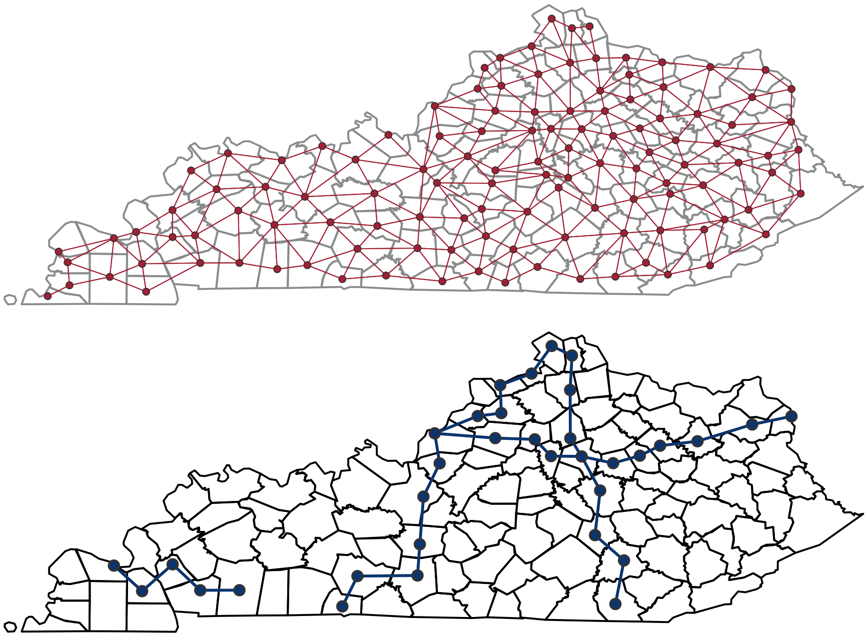


Figure 10: Adjacency between Kentucky counties (top) and observed highway paths (bottom).

Interstate period. Geographic distributions of the covariates are depicted in figure 11.

Parameter estimates are plotted in figure 12. Parameter signs for stated decision criteria (distance, population, and agriculture) are consistent with qualitative evidence, though statistically insignificant. The effect of population are perhaps understated as a result of Lexington and Louisville being held fixed, but relaxing this assumption would make the model less plausible. Coefficients may be interpreted in relation to one another: for example, the coefficient of county representation, β_{plan} is more than double that of log-population. This implies that with respect to the probability of highway construction, holding all other nodes fixed, the marginal effect of political representation in the highway decision-making process is roughly equivalent to a sixfold increase—an e^2-fold multiple—in total population. Note that in the RPM context, interpretation of coefficients with reference to "an otherwise identical" unit implies that the reference unit is not only identical on covariates, but also on its position in the graph—which is to say, the "otherwise identical" unit must essentially be the same unit.

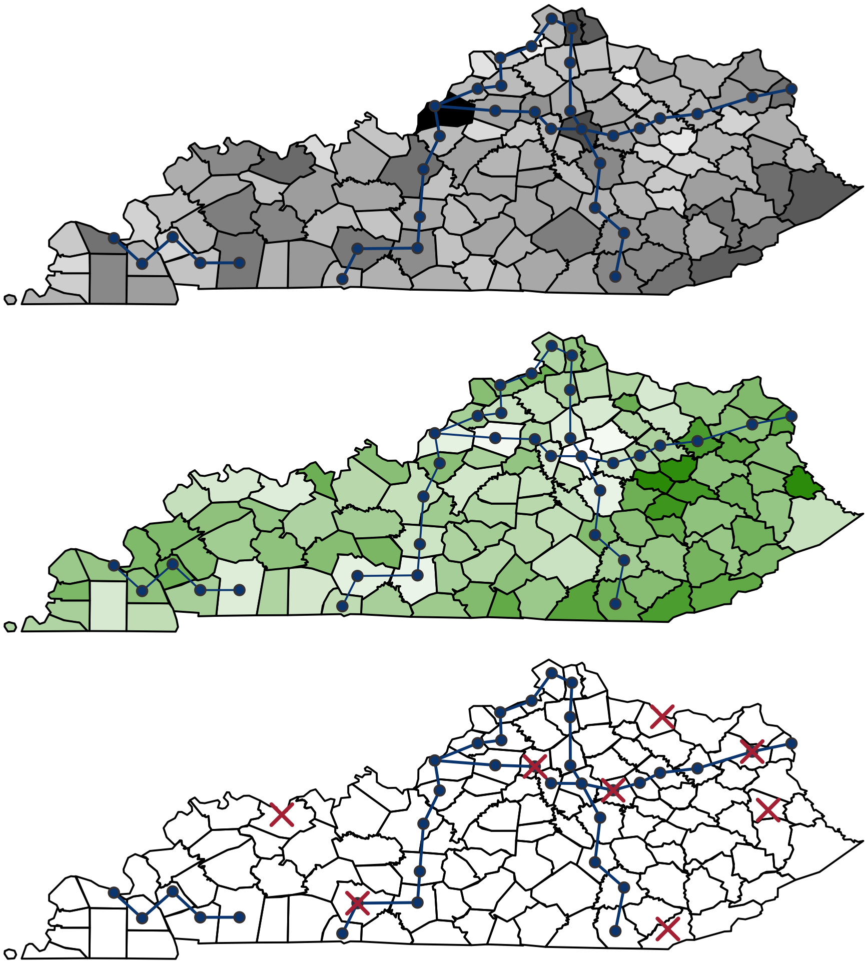


Figure 11: Log population of Kentucky counties (grey, top); total value of agricultural products (green, center); and presence of a hometown member on the Postwar Planning Committee ("×," bottom).

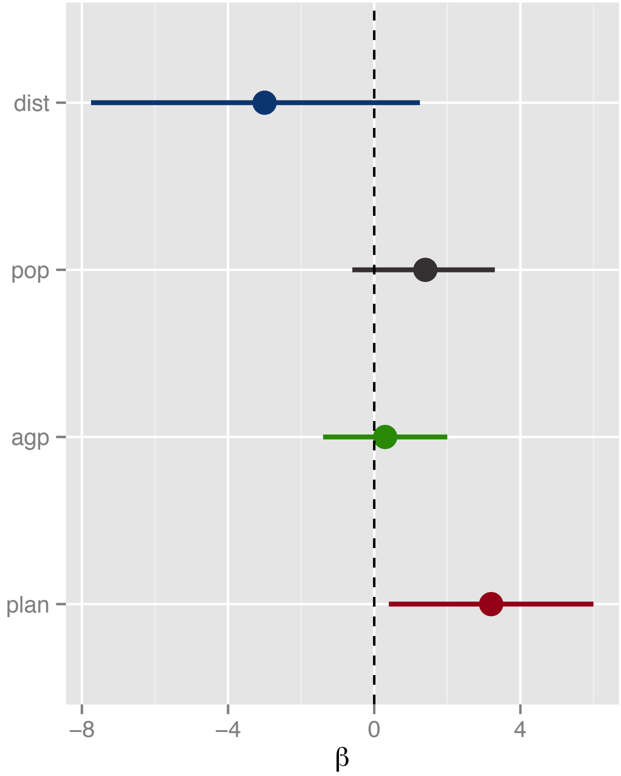


Figure 12: 95% posterior credible intervals. RPM coefficient estimates do not have an immediate substantive interpretation, except in relation to one another. Positive coefficients indicate that an increase in the corresponding covariate in county i would increase the probability of highway construction, holding all other counties constant. The sign on distance is negative and signs on both log-population and log-agricultural product are positive, as expected. The coefficient on county participation in the state Planning Committee is significant and surprisingly large; it is estimated as equivalent to more than a sixfold increase in population.

5 Concluding Remarks and Future Directions

In this paper, I demonstrate that paths can be effectively modeled and develop tools that allow social scientists to assess a range of previously untestable hypotheses. I show that while dependence between observations presents a statistical challenge, it does not prevent inference on factors shaping the trajectories of paths. This opens the door for analyses on a vast range of unexploited data sources.

For example, an ongoing extension uses the RPM to draw inferences about the causal effects of paths. In contrast to current approaches that require untenable assumptions about spatial dependence in the outcome variable (Cox, 1958; Rubin, 1980), the proposed approach requires only an assumption about the treatment-assignment process (Rubin, 1991). Such an approach is particularly attractive when interference between units is possible (Hudgens and Halloran, 2008), as seems likely for, e.g., connective infrastructure. Causal effects can be estimated using individual exposure probabilities (Aronow and Samii, n.d.), simulated by RPM draws with the estimated parameters; uncertainty about the treatment-assignment process can easily be accounted for by sampling parameters from the RPM posterior and, for each parameter draw, incorporating the estimated variance of Aronow and Samii's procedure via simulated joint exposure probabilities. In addition, focusing attention on the treatment-assignment process allows theoretically motivated hypotheses about specific forms of interference—e.g., spillover effects on counties near highways, reinforcement effects when trading partners are simultaneously connected—to be formalized and tested as in Bowers, Fredrickson and Panagopoulos (2013).

Much work remains to be done on the random-path model, particularly with respect to model specification. Even when the path-generating process is known to depend only on a few factors, as in the Kentucky example, qualitative theory provides little guidance about how those variables should be operationalized in the RPM context. Moreover, the bias induced by omitted variables is unknown. Adaptations of sensitivity analyses and test statistics (for the equivalent of "examining the residuals”) are areas of ongoing research.

A Appendix

A.1 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.1.1 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 $\tilde{G}_\mathcal{T}(G)$ 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.1.2 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." (For convenience, the LERW procedure is restated in algorithm 4 in slightly more general form.) 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

$$\begin{aligned} \Pr\left(\texttt{LERW}(\tilde{G}, u_2, u_1) = \psi\right) &= \sum_{T \in \mathcal{T}} \Pr\left(\texttt{LERW}(\tilde{G}, u_2, u_1) = \psi \mid W = T\right) \Pr(W = T) \\ &= \sum_{T \in \mathcal{T}} \mathbf{1}\{\psi \text{ is a subgraph of } T\} \frac{1}{|\mathcal{T}|}. \end{aligned} \tag{4}$$

□

A.1.3 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

```
Procedure LERW_multiple($\tilde{G}, \gamma_0$, stopping set)
    initialize $\psi = (\gamma_0)$,   $i = \gamma_0$
    while $i \notin$ stopping set do
        sample $j$ uniformly from $\mathcal{N}_i$
        step to $i = j$ and append to $\psi$
    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 $\psi$
        end
        $t$+=1
    end
    return $\psi$
```

Algorithm 4: Generic form of the loop-erased random walk that stops when any node in the stopping set is encountered.

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.[8] Thus, $\tau(\tilde{G}) = \tau(\tilde{G} - e) + \tau(\tilde{G}/e)$.

A.1.4 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,

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

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

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 4 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 likelihood ratio of the latter relative to the former, which is simply

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

□

A.2 Illustration of Proposition 1

Consider the wheel graph in figure 13. The first row of figure 14 shows three possible paths from node A to node E (solid lines). Below each depicted path, spanning trees are formed by adding edges (dotted lines) so that all nodes are connected without cycles. For the (A, B, E) path, there are eleven unique spanning trees that can be generated in this way; thus, (A, B, E) is eleven times more likely to be drawn by a LERW relative to the (A, B, C, D, E) path. Note that (A, B, C, D, E) is already a spanning tree, so no edges can be added without forming a cycle.

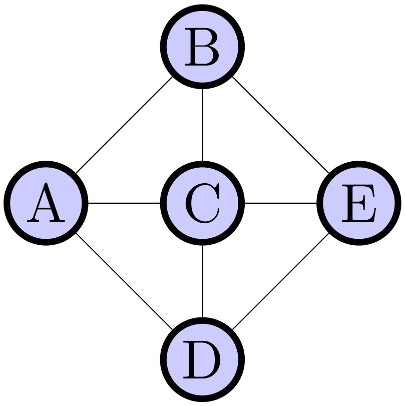


Figure 13: A wheel graph with five nodes.

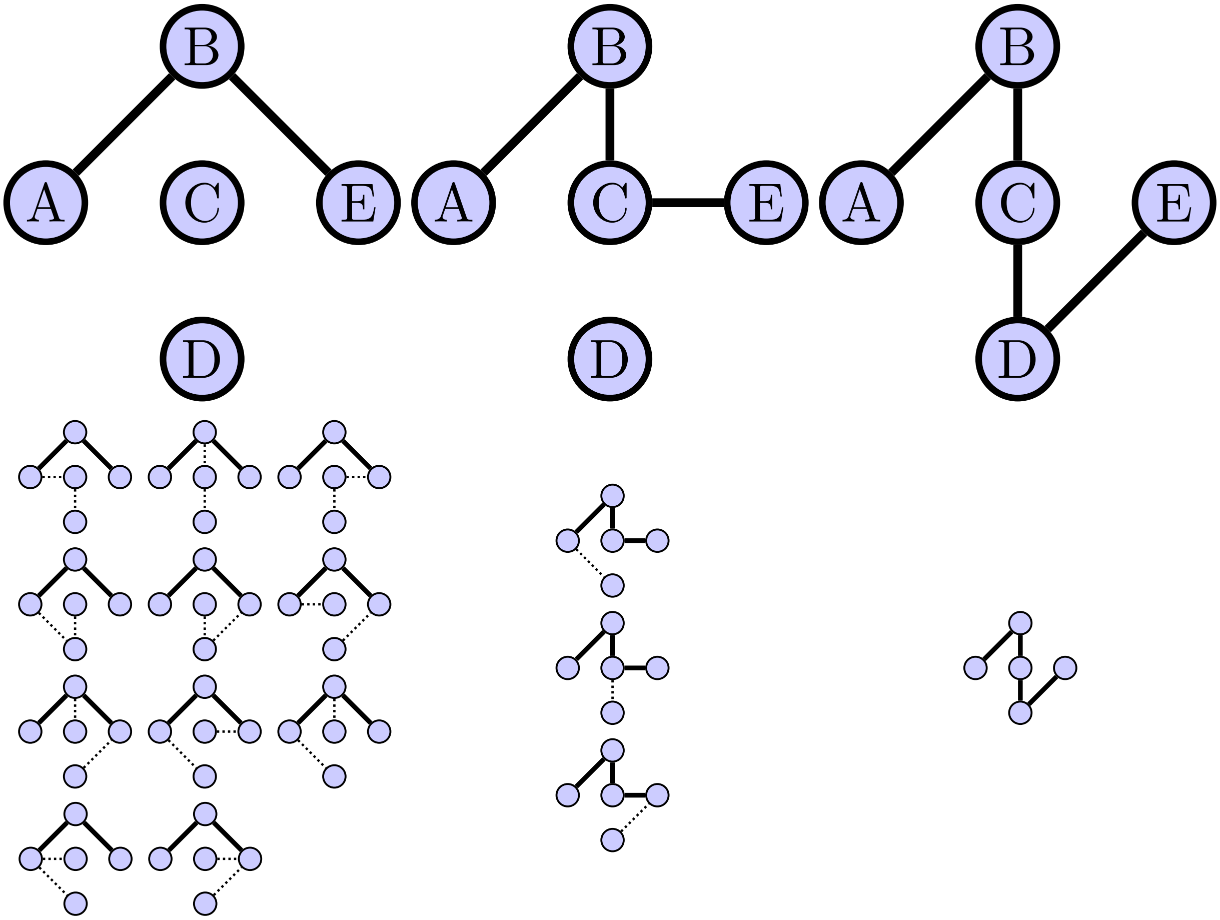
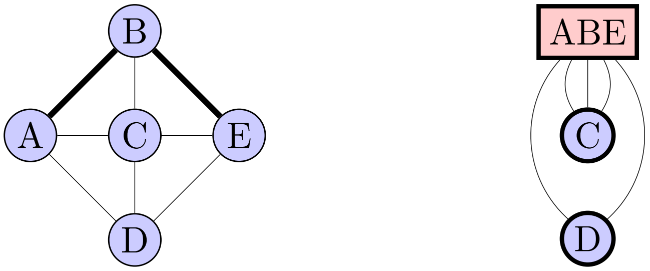


Figure 14: The top panel depicts three possible paths from A to E, with solid edges. The LEWR samples each path with probability proportional to the number of spanning trees that can be constructed from that path, shown in the bottom panel beneath each path (additional edges need to construct a spanning tree are shown with dotted lines). Thus, (A, B, E) is eleven times more likely to be drawn by a LERW from A to E, relative to (A, B, C, D, E).

The deletion-contraction recurrence can be used to enumerate spanning trees by recursively dividing them into those that do not use a particular edge (the set of all spanning trees on the deletion) and those that do (spanning trees on the contraction). The deletion of an edge e from G, denoted $G \backslash e$, removes e from a graph. The contraction operation, $\tilde{G}/e$, involves fusing e's endpoints into a single new node, then removing the self-loop on the new fused node. The contracted graph is permitted to retain multiple edges from the fused node to others. For example, in figure 15, the graph is contracted along the (A, B, E) path by eliminating the A–B and B–E edges, then fusing nodes A, B, E into a new node, ABE. The resulting multigraph contains three edges from ABE to C, corresponding to the former edges A–C, B–C, and E–C. The adjacency and degree matrices of the full graph (A and D) and contracted graph (A^* and D^*) are shown directly beneath.



$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix} \qquad A^* = \begin{bmatrix} 0 & 3 & 2 \\ 3 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

$$D = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix} \qquad D^* = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Figure 15: The (A, B, E) path on a wheel graph, W, is shown at top left. At top right, contraction along the A–B and B–C edges. Corresponding adjacency and degree matrices of each given below.

Then, it is straightforward to confirm that in this example, det $L^*_{i,j} = \det(D^* - A^*)_{i,j} = 11$ for any i and j, where $L^*_{i,j}$ denotes the minor of L^* formed by removing the ith row and jth column. This is precisely the number of trees that were explicitly enumerated in figure 14.

References

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

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.

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.

Burgess, Robin, Remi Jedwab, Edward Miguel, Ameet Morjaria and Gerard Padr i Miquel. 2013. The Value of Democracy: Evidence from Road Building in Kenya. Working Paper 19398 National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w19398*

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

Cox, D. R. 1958. *Planning of Experiments.* New York: Wiley.

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

Dormann, Carsten F., Jana M. McPherson, Miguel B. Arajo, Roger Bivand, Janine Bolliger, Gudrun Carl, Richard G. Davies, Alexandre Hirzel, Walter Jetz, W. Daniel Kissling, Ingolf Khn, Ralf Ohlemller, Pedro R. Peres-Neto, Bjrn Reineking, Boris Schrder, Frank M. Schurr and Robert Wilson. 2007. "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review." *Ecography* 30(5):609–628.

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.

Frejinger, E., M. Bierlaire and M. Ben-Akiva. 2009. "Sampling of alternatives for route choice modeling." *Transportation Research Part B: Methodological* 43(10):984–994.

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.

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

Hudgens, Michael G. and M. Elizabeth Halloran. 2008. "Toward Causal Inference With Interference." *Journal of the American Statistical Association* 103(482):832–842.

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

Mai, Tien, Emma Frejinger and Mogens Fosgerau. n.d. A nested recursive logit model for route choice analysis. Technical Report 63161. http://mpra.ub.uni-muenchen.de/63161/.

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

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.

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.

Rubin, Donald B. 1980. "Discussion of "Randomization Analysis of Experimental Data in the Fisher Randomization Test," by Debabrata Basu." *Journal of the American Statistical Association* 75(371):575–582.

Rubin, Donald B. 1984. "Bayesianly justifiable and relevant frequency calculations for the applied statistician." *Annals of Statistics* 12:1151–1172.

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.

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.

U.S. National Interregional Highway Committee. 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.

U.S. Public Roads Administration. 1939. Message from the President of the United States transmitting a letter from the secretary of agriculture, concurred in by the secretary of war, enclosing a report of the Bureau of Public Roads, United States Department of Agriculture, on the feasibility of a system of transcontinental toll roads and a master plan for free highway development. 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.

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.