WRITING SAMPLES

1 Rooted America: Immobility and Segregation of the Inter-county Migration Networks

Dissertation chapter; conditionally accepted by *American Sociological Review*

2 California Exodus? A Network Model of Population Redistribution in the United States

Dissertation chapter; published in *Journal of Mathematical Sociology*

3 Geographical Patterns of Social Cohesion Drive Disparities in Early COVID Infection Hazard

Published in *Proceedings of the National Academy of Sciences (PNAS)*

4 Spatial Heterogeneity Can Lead to Substantial Local Variations in COVID-19 Timing and Severity

Published in *Proceedings of the National Academy of Sciences (PNAS)*

5 Parameter Estimation Procedures for Exponential-Family Random Graph Models on Count-Valued Networks: A Comparative Simulation Study

Dissertation chapter; published in *Social Networks*

6 Marginal-preserving Imputation of Three-way Array Data in Nested Structures, with Application to Small Areal Units

Published in *Sociological Methodology*

Peng Huang
Department of Sociology
University of California, Irvine
peng.huang@uci.edu

Rooted America: Immobility and Segregation of the Inter-county Migration Networks

Abstract

Despite the popular narrative that the United States is a "land of mobility," the country may have become a "rooted America" after a decades-long decline in migration rates. This paper interrogates the lingering question about the social forces that limit migration, with the empirical focus on the internal migration in the United States. We propose a systemic, network model of migration flows, combining demographic, economic, political, geographical factors and network dependence structures that reflect the internal dynamics of migration systems. Using valued temporal exponential-family random graph models, we model the network of inter-county migration flows during the 2011-2015 period. Our analysis reveals a pattern of *segmented immobility*, where fewer people migrate between counties with dissimilar political contexts, levels of urbanization, and racial compositions. Probing our model using "knockout experiments" suggests that one would have observed approximately 4.6 million (27%) more inter-county migrants each year were the segmented immobility mechanisms inoperative. The paper offers a systemic view of internal migration and reveals the social and political cleavages that underlie geographical immobility in America.

Keywords

migration, social networks, political polarization, immobility, segregation

While the active drivers of migration have been extensively studied, there has been less attention to the factors that *hinder* migration – a research gap that has been called the "mobility bias" within the migration literature (Schewel 2020). The relatively overlooked phenomenon of immobility is important in its own right, having substantial consequences for the social world. As migration influences the functioning of labor market (Hyatt et al. 2018), the landscape of stratification and social mobility (Jasso 2011), and the sociocultural meanings in everyday lives (Bauman 2000; Mata-Codesal 2015), mechanisms that impede migration can have outcomes that extend far beyond the migration system itself.

Understanding immobility is an especially apt challenge in the context of the modern United States. Long thought of as a "rootless society" (Fischer 2002) with high geographical mobility (Long 1991; Steinbeck 1939), the U.S has arguably turned into a "rooted America" after a decades-long decline in migration rates (DeWaard et al. 2020; Frey 2009). While the reality of low migration rates is clear, explanations for current population immobility are less well-developed. Macroeconomic studies have so far found that demographic and socioeconomic structures are not sufficient to explain observed levels of immobility, and neither are the business composition of labor market nor properties of the housing market (Hyatt and Spletzer 2013; Hyatt et al. 2018; Molloy et al. 2011, 2017). A broader sociological view suggests the potential for cultural, political, and other social forces as possible explanatory factors (Tiebout, 1956; Massey and Denton, 1993; Gimpel and Hui, 2015; Stockdale and Haartsen, 2018). Moreover, the migration system has its own intrinsic feedback mechanisms that could endogenously sustain or undermine further migration (Bakewell, 2014; de Haas, 2010), which may also play a role in the population immobility. Probing the combined influence of these myriad factors requires a *systemic* treatment of the U.S. internal migration, allowing us to simultaneously examine the joint impact of social, economic, political, and demographic mechanisms on flows of migrants throughout the country. This paper pursues such an analysis, with the objective of identifying the factors associated with both mobility and immobility in contemporary America.

Broadly, extant research on drivers of U.S. migration and immobility shares two characteristics.

First, most research examines migration from an economic perspective, assuming that most, if not all, migration is *labor migration*, driven by economic incentives.[1] Yet, decisions regarding residential settlement are not purely economic (Ryo 2013): political climate, racial composition, and urbanization of local communities are potential contributors to the phenomenon (Brown and Enos 2021; Cramer 2016; Massey and Tannen 2018). This paper incorporates the sociocultural and political perspectives into the analysis of U.S. immobility.

A second dominant characteristic of the extant literature on U.S. migration is an approach that treats migration as a feature of geographical areas, examining the correlates between net migration rates into or out of states or counties and their demographic or economic characteristics. Although convenient, this practice of reducing the interconnected migration system into local features of areal units introduces two limitations. First, by aggregating across origins and destinations for migrants emigrating from or immigrating into a given area, it obscures the *interactive effects* from the sending and receiving areas, such as their political or cultural similarity and differences in employment rates. Second, it does not allow for treatment of the *internal dynamics* of the migration system (de Haas 2010), in particular the presence of mechanisms such as return or stepwise migration, where the flow of migrants from one place to another can in turn affect the flow of migrants from that destination to others. Since migration is a relational process between places of origin and destination, and migration flows can influence each other, this paper takes a systemic, network approach that shifts analysis from the migration rates of areal units to the migration flows *between* areal units. By leveraging migration systems theory and social network methods, we show that dissimilarities between counties are important contributors to the immobility of American society.

To advance our understanding of the social forces behind geographical immobility in modern America, we here adopt a comprehensive theoretical framework incorporating geographical, demographic, economic, political, and social influences on migration and perform a systemic analysis of internal migration as an evolving valued network of migration flows.[2] Using valued temporal exponential-family random graph models (valued TERGMs), this paper analyzes the network of intercounty migration flows of the United States from 2011 to 2015. We identify a pattern of

segmented immobility, where, net of other factors, less migration happens between counties with dissimilar political contexts, levels of urbanization, and racial compositions. We probe this mechanism using an *in silico* "knockout experiment," which suggests that in a counterfactual world without segmented immobility (but holding all other factors constant), we would expect to have seen approximately 4.6 million (27%) more intercounty migrants in the United States each year. This implies that social and political cleavages in America are substantial contributors to immobility, and potentially exacerbate growing trends towards geographical segregation. Further, we also examine the relationship between internal and international migration flows, showing that - contrary to the balkanization thesis (Frey 1995a,b) - international migration into a county is positively associated with its overall domestic mobility, and does not promote net outflows of residents. The model also identifies the internal dynamics of migration systems (de Haas 2010), including a suppression of what we dub "waypoint" flows (i.e., balanced in- and out-flows of an areal unit) alongside strong patterns of reciprocity and perpetuation. While the data availability constraints us to focus on understanding population immobility in the 2010s, the empirical evidence together with our proposed theoretical and methodological frameworks opens the door to unpack the long-term phenomenon of population immobility. This paper thus joins the growing literature that grapples with the mobility bias in migration studies (Schewel 2020), demonstrating how a comprehensive analytical framework and a systemic, network approach offers new insights about immobility, and more broadly, the dynamics of population movement among social and geographical spaces.

THEORY

Existing literature defines immobility as "continuity in an individual's place of residence over a period of time" (Schewel 2020:344). Since immobility is not only an individualistic phenomenon, but also a population and social one, here, we offer a macrosociological definition of immobility, which is a lack of population exchange between localities. Drivers of immobility, in terms of this framework, are defined as factors that *reduce* migration rates relative to what would be expected

in their absence. The scarcity of migration in an immobile society has substantial impacts. Since migration is a critical channel for people to respond to fluctuations of local economy, population immobility implies a rigid labor market with lower productivity, higher unemployment rate, and more prolonged recession when experiencing economic shocks (Hyatt et al. 2018). Moreover, migration also serves as a way of improving life chances (Jasso 2011; Weber 1922) and coping with adverse events (Spring et al. 2021). Population immobility thus has important ramifications for social mobility, stratification, and poverty (Briggs et al. 2010; Clark 2008; Jasso 2011).

Immobility is not merely the flip side of mobility, but carries its own sociocultural meanings. As the aspiration-ability model argues, migration requires both aspiration to migrate and the ability to realize that aspiration (Carling and Schewel 2018). This means that immobility is not necessarily a passive outcome of simply staying in place, but can be a conscious choice to remain. In line with this view, recent literature has begun augmenting the widely discussed notion of "cultures of migration" with the notion of "cultures of staying" that facilitate and maintain immobility (Stockdale and Haartsen 2018). The level of population (im)mobility can in turn impact the broader social norms of a society; a mobile society may have a prevailing nomadic culture, while the dominant culture of an immobile society may be sedentary (Bauman 2000; Mata-Codesal 2015).

Understanding immobility is especially relevant in the American case. From the earliest observations of Tocqueville (1834) and Ravenstein (1885) to Steinbeck (1939), America has long been considered a "restless" or "rootless" society with high geographical mobility. Yet, after a decades-long decline in its migration rate, the contemporary America has arguably become a "rooted" society with considerable population immobility. However, as Herting et al. (1997:267) have noted, sociological research on U.S. mobility has "narrowed and now focused almost exclusively on mobility of a purely economic or occupational variety," with much less focus on mobility across geographic space. In migration studies, research has been historically focused on studying the social forces that lead to migration, but largely neglected the counter forces that *inhibit* people from moving, a tendency that Schewel (2020) described as the "mobility bias." A lack of research on geographical mobility in American sociology, together with the scarcity of theoretical and em-

pirical work on immobility in migration studies, has led to gap in our knowledge regarding the mechanisms behind population immobility in contemporary American society.

Culture and Politics of Immobility

While the immobility of the U.S. population has received less sociological attention, economists and geographers have conducted empirical analyses on this matter (e.g., Cooke 2013; Jia et al. 2022; Kaplan and Schulhofer-Wohl 2017; Treyz et al. 1993). These studies have identified important connections between the labor market and migration rates, but their findings largely rely on the assumption that most, if not all, migration is *labor migration*, driven by economic incentives. The economic perspective has a fundamental role in explaining migration and immobility; the relative gains in moving, and costs associated with both transaction costs and losses of specialized local investments *are* factors that shape migration. But there also exist other factors, such as regionally specific cultural values and locally conventional ways of understanding opportunity (Carling 2002; Carling and Schewel 2018), as well as preferences for particular local policies or political regimes (Tiebout 1956). Indeed, recent research on American economy has shown that over the past several decades, migration has not been effective in responding to fluctuations and shocks in labor markets (Dao et al. 2017; Jia et al. 2022). Relatedly, macroeconomic factors have not been found to have a strong correlation with migration rates in the U.S. (Hyatt et al. 2018; Hyatt and Spletzer 2013; Molloy et al. 2017). Therefore, while economic forces are important ingredients in a viable model of the migration system, a comprehensive analysis of immobility demands considerations of other social institutions.

Although thinking on internal migration in the large has been dominated by labor market considerations, sociologists have given considerable attention to other factors when studying migration at smaller scales (e.g., across neighborhoods). For instance, research on residential segregation has long identified how people with different racial identities and political beliefs become segregated from each other (Bishop and Cushing 2009; Krysan and Crowder 2017; Massey and Denton 1993), including the accumulated influence of even relatively weak preferences for same-group interaction

(Schelling 1969; Sakoda 1971); the latter can act as a powerful macro-level sorting force, even in the presence of economic or other factors (e.g., Butts 2007). While much of this work has focused on racial segregation, more recent work has also probed segregation along political or cultural axes. For instance, Brown and Enos (2021) found that a large proportion of American adults live in neighborhoods where most residents share the same partisanship. Gimpel and Hui (2015) used a survey experiment to show that people evaluate more favorably towards properties in areas with predominantly co-partisan neighborhoods. As social cleavages might deter people from settling in places with distinct identities and beliefs, the social gaps between rural and urban areas and those among different parts of the continent such as the South and the coastal regions (Cramer 2016; Hochschild 2018), may also contribute to the inhibition of geographical movement. At another scale, in the contexts of international migration, migration studies have long stressed the roles of cultures and politics in shaping population mobility (Castles et al. 2013; Cohen and Sirkeci 2011; Jennissen 2007; Massey et al. 1999; Vögtle and Windzio 2022; Waldinger and Fitzgerald 2004). Following this thread, this paper incorporates the political, racial and rural-urban structures in investigating American immobility.

Systemic Theories of Migration

The second characteristic of the extant literature on U.S. immobility is that studies usually view migration as a feature of geographical areas. This approach examines the characteristics of an areal unit that influence its net immigration and emigration rates, such as percentages of current residents who are immigrants and/or emigrants. It is in essence a marginal approach that sums up (i.e., marginalizes) the migration flows from/to each areal unit across all destinations/origins to describe the overall mobility of each place. The marginal approach is empirically straightforward, and has unquestionably contributed to our understanding regarding the driving forces of migration by identifying the associations between demographic and economic features of an areal unit and the scale of its population inflows or outflows (e.g., Partridge et al. 2012; Treyz et al. 1993). Yet, migration - by definition, population moving from one place to another - is *inherently* relational,

having properties that cannot be reduced to the features of individual areal units. For instance, studies considering net in- or out-migration rates in isolation must choose either the sending or receiving area as focus of analysis (thereby obscuring the joint roles of areas as origins and destinations), or must merge in- and out- migration to obtain a net migration rate (which confounds inflows and outflows). Beyond the fact that every pairwise migration flow among sending and receiving areas depends on both the properties of the sender and the receiver, such studies are unable to account for relational factors, such as geographical proximity and political difference between areal units. Neither can this approach consider the interactions among migration flows, such as reciprocal population exchange ($A \rightarrow B$ & $B \rightarrow A$) arising from return migration. Probing such mechanisms requires a different theorization of the migration process, a systems theory of migration.

Such systemic thinking has a long tradition in migration studies under the umbrella of migration systems theory (MST, Bakewell 2014; Fawcett 1989; Kritz et al. 1992; Mabogunje 1970; Massey et al. 1999). A comprehensive theory that concerns the complex interactions among various elements related to migration, such as flows of people, information, (formal and informal) institutions, and strategies (Bakewell 2014), MST identifies *interconnectivity* as a key feature of migration. As de Haas (2010:1593) summarized, a migration system is "a set of places linked by flows and counter-flows of people, goods, services and information, which tend to facilitate further exchange, including migration, between the places." The theoretical focus on flows *between* origin and destination suggests a relational analysis of migration, integrating push and pull factors in one single analytical framework (Lee 1966). Fawcett (1989) demonstrates this with a theoretical framework of "linkages" in MST, focusing on how various linkages between origin and destination shape the migration in between. Among the linkages Fawcett (1989:677) discusses, here we focus on the relational linkages, "derived from comparison of two places." Instead of studying how a state's or a county's political climate influences its net marginal migration rate (e.g., Charyyev and Gunes 2019; Preuhs 1999), an analysis of relational linkages examines how the *difference* in political climates between counties influences the number of people migrating from one to the

other.

Another critical implication from the interconnectivity feature of migration systems is the presence of internal dynamics of migration (Bakewell et al. 2016b; de Haas 2010; Mabogunje 1970). As Mabogunje (1970:16) put it, the migration system is "a circular, interdependent, progressively complex, and self-modifying system in which the effect of changes in one part can be traced through the whole of the system." Similarly, Fawcett (1989:673) argued that the migration systems framework "brings into focus the interconnectedness of the system, in which one part is sensitive to changes in other parts." This means that migration is not a pure product of exogenous social forces. It forms a system with endogenous processes, where one migration flow can promote or suppress another migration flow. For example, since migrants transmit information and social connections when they move, the migration flow from Arizona to Texas brings job information and personal contacts along, potentially inspiring migration in the opposite direction. Internal dynamics like this can lead to an endogenous accumulation of migration net of exogenous social and economic influences.

Migration Systems Through a Network Lens

The insight of interconnecitivity from MST resonates with that of social network analysis. Indeed, past research has employed social network analysis to study migration systems (Charyyev and Gunes 2019; Desmarais and Cranmer 2012; DeWaard et al. 2012; DeWaard and Ha 2019; DeWaard et al. 2020; Hauer 2017; Leal 2021; Liu et al. 2019; Nogle 1994; Vögtle and Windzio 2022; Windzio 2018; Windzio et al. 2019). This school of MST, called by Bakewell (2014) the "abstract system," interrogates the macro-level migration patterns by analyzing migration networks consisting of localities (in network terms, *nodes*) and migration flows between each directed pairs of localities (in network terms, *edges*).[3] Network analysis effectively captures the two critical implications of MST, relational linkages and internal dynamics of migration systems, bringing new perspectives compared to the marginal approach of migration, commonly employed in studies of U.S. immobility. Rather than viewing localities/places as units of analysis, the network approach

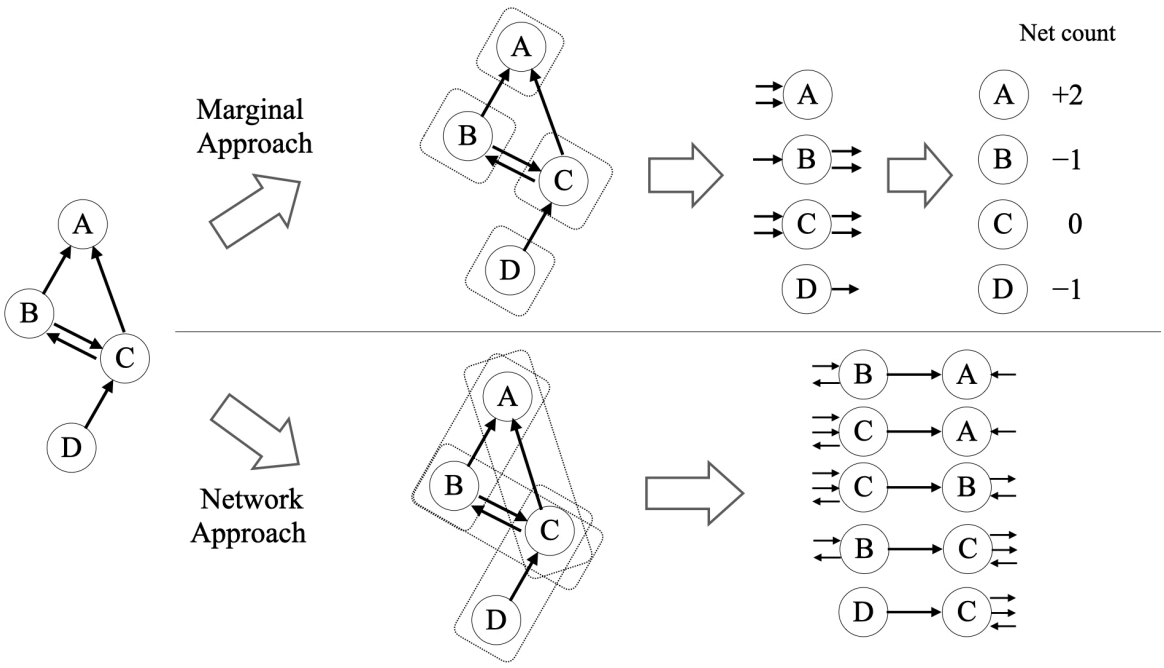


Figure 1. Schematic illustration of the marginal approach versus the network approach
Note: The marginal approach takes geographical areas as units of analysis, and tends to condense the in- and out- migration flows into a single number about net migration rate/count of a geographical area. The network approach takes each migration flow between a directed pair of geographical areas as an analytical unit. This approach incorporates origin and destination in understanding their joint influence on migration flows; it also preserves the local structural properties of migration flows, allowing systemic patterns to be examined.

takes migration flows between places as analytical units. This perspective preserves information regarding emigration and immigration processes, enabling analysis of how characteristics of origin and destination *interact* to influence migration flows, a relational account of linkages in migration systems. The network approach also examines the internal dynamics of migration systems, by studying the dependence structure among migration flows. The dependence structure identifies how migration flows are associated with each other, net of the exogenous contexts such as the economic and political environments. Taking the above-mentioned example of reciprocity, the network approach measures whether and to what extent, an increase of one migration flow (e.g., Los Angeles to Baltimore) is associated an increase in its opposite flow (Baltimore to Los Angeles), net of other factors. The dependence structure can further go beyond a pair of places and describes how the whole network system of migration flows are interconnected, such as how the

migration inflows of Denver are associated with its outflows, which in turn serve as the inflows of another places, Dallas, Atlanta, etc. Figure 1 illustrates the network approach in contrast to the marginal approach.

While the network approach introduces unique perspectives overlooked by the marginal approach, its insights has not yet been fully appreciated. One notable characteristic of prior research on migration networks is the focus on the "diversity" rather than the "intensity" of migration flows (DeWaard and Ha 2019; Leal 2021; Vögtle and Windzio 2022; Windzio 2018; Windzio et al. 2019). In other words, extant research examines the *number* of migration flows rather than their *magnitudes*. This is associated with the practice of dichotomizing migration flows into two statuses, either no migrants versus at least one migrant, or few migrants versus many migrants (though Windzio (2018) and Windzio et al. (2019) divide them into more (5) statuses in some parts of their research). This approach is compatible with the common practice in social network research of approximating social relations by a binary form, facilitating the use of existing network theories and methods to describe the migration system. While analyzing the "diversity" of migration flows offers useful knowledge about the migration system, it ignores the rich information about the variation in migration magnitudes. The intensity of migration flows becomes a critical question when it comes to understanding population immobility. In particular, DeWaard et al. (2020) find that the decline of U.S. migration is not due to the decline in the diversity of migration flows (the number of county pairs with population exchange), but the decline in the intensity of migration flows (their average count of migrants). Studying the intensity of migration flows requires describing migration networks in a valued form, where the edges are not binary, but take quantitative values. Since the quantitative feature of migration intensity is critical in grappling with the question of population immobility, this paper bridges migration systems theory and recent advances in statistical and computational methods for valued network analysis (Huang and Butts 2024; Krivitsky 2012). We formally theorize the relational linkages and internal dynamics in the expressions of valued networks, developing a roadmap to quantitatively describe and test the interconnectivity of population flows.

On the side of migration systems, new theoretical insights are needed for studies of immobility. MST is not an exception from the mobility bias critique of migration theories (Schewel 2020). As de Haas (2010) argues, MST has historically focused extensively on migration-facilitating mechanisms that lead to the perpetuation of migration flows, but largely overlooked the migration-undermining mechanisms that lead to the decline of migration flows. Building on this critique, a line of theoretical and empirical research studies why some instances of pioneer migration drive the formation of migration systems while others do not, and the endogenous mechanisms that can undermine the migration system (Bakewell et al. 2012, 2016a; de Haas 2010). Bakewell et al. (2016a) further go beyond the MST framework, as they pursue the notion of incorporating scenarios where the migration systems fail to form or perpetuate. Unquestionably, this is a promising direction to further the theorization of migration dynamics. Yet, for our focus of internal migration in the contemporary U.S., the migration system has been in existence for generations, and is unlikely to vanish in the near future. Therefore, the migration system is still a useful research subject and perspective, where we explore the social mechanisms that immobilize population from migrating.

The network approach inspires us to consider population immobility from a relational perspective. We conceptualize the pattern of *segmented immobility*, that in a society where people cluster in geographical segments based on their cultural and political traits, immobility can occur due to people's tendency to avoid migrating towards places with divergent environments. By jointly incorporating origin and destination in an analytical framework, the relational perspective allows us to examine the influence of dissimilarity between counties on the magnitude of migrant populations moving between them, connecting population immobility with segregation and polarization. Apart from examining the pattern via a hypothesis testing lens, we further utilize the idea of "knockout experiments" broadly employed in the experimental sciences to directly quantify its contribution to immobility. Originating in biomedical research, a knockout experiment probes the functional role of a system component by removing or inactivating it, comparing normal system behavior with behavior when the component is "knocked out" (Hall et al. 2009; Vogel 2007). In

social sciences, knockout experiments are performed *in silico*, where researchers simulate the potential social outcomes when certain social forces were removed. The knockout experiment can be considered as a model-based thought experiment (*Gedankenexperiment*, Einstein et al. 1935), in which we predict the social outcomes of interest under a counterfactual scenario where certain social effects are inoperative. In our case, we compare the total number of migrants observed in the real world to that simulated when segmented immobility mechanisms are knocked out. This theoretical exercise allows us to leverage the power of modern, generative network models to gain insights into the functioning of the migration systems.

HYPOTHESES

Relational Linkages: Political Segregation and Segmented Immobility

Decisions about migration, a behavior aiming at improving life chances (Jasso 2011), typically come out of a comparison between place of departure and destination. Moving from current place of residence, will the destination be adaptive? One critical dimension in drawing an answer is the political environment of the origin and the putative destination communities. Rising political polarization has divided Americans along the party lines (Levendusky 2009), where social cleavage by political ideology extends to a growing array of public opinions (Baldassarri and Gelman 2008; DellaPosta 2020) and choice of lifestyles (DellaPosta et al. 2015), and has lead to segregated social networks and tensions in relationships such as familial interactions (Chen and Rohla 2018; DiPrete et al. 2011). This political alignment also happens across space, with distinct political consciousness across geographical regions, rural and urban lands, and local neighborhoods (Bishop and Cushing 2009; Cramer 2016; Hochschild 2018). Recent spatial analysis on partisan isolation reveals that a large fraction of American adults lives in places where almost no one in their neighborhood votes in a manner opposed to their own (Brown and Enos 2021). They also found that this pattern is prevalent nationwide and is a distinct pattern from segregation in other dimensions such as across racial lines. This state of affairs is also overtly recognized within American polit-

ical discourse, where media outlets routinely make distinctions between "red" (conservative) and "blue" (liberal) regions, and ascribe (correctly or not) a large body of cultural and political traits to both the regions and their inhabitants (Badger et al. 2018; Wallace and Karra 2020). To the extent that individuals are likely to both affiliate with the political culture of their area, and regard their opposites on the political spectrum with suspicion and even hostility (Iyengar et al. 2012, 2019), people may be unwilling to migrate between regions with differing political cultures. Even setting aside motivations arising from political culture, according to the public choice theory and the consumer-voter model, people should still be more willing to migrate to regions whose governments most closely match their own policy preferences (Dye 1990; Tiebout 1956), with those from solidly "red" areas preferring to move to other "red" areas, and likewise for those from "blue" areas. Empirical analyses using various data and methods generally confirm the existence of migration preference towards co-partisanship (Tam Cho et al. 2013; Gimpel and Hui 2015; Liu et al. 2019), though with some counter evidence (Mummolo and Nall 2016). Together they motivate the following hypothesis:

Hypothesis 1.1: Ceteris paribus, *the more dissimilar counties are in their average political orientation, the lower the migration flow between them.*

The limited population exchange between geographical segments with dissimilar social environments, or what we call *segmented immobility,* may not be unique to the political dimension, but would rather be a pervasive pattern arising from people's evaluation of places along multiple dimensions. One of the underlying mechanisms that can lead to such a pattern is homophily. Homophily refers to people's tendency to be connected to and interact with those similar to themselves in various characters such as racial and ethnic identity, religious belief, political ideology, personality, and normative inclination like altruism (DiPrete et al. 2011; Leszczensky and Pink 2019; McPherson et al. 2001; Moody 2001; Smith et al. 2014; Wilson et al. 2009). Homophily occurs not only within personal networks, but is also a spatial phenomenon, where people tend to live close to others with similar racial identity, economic background, or political ideas (Bishop and Cushing 2009; Massey and Denton 1993; Intrator et al. 2016). A social process that can give

rise to this spatial pattern is that residents choose to migrate towards places where people similar to them concentrate, but avoid destinations with identities different from their own (Crowder et al. 2012; Massey et al. 1994b; Schelling 1969). Although literature about this residential sorting process focuses primarily on mobility among neighborhoods in urban areas, we argue that a similar process may also work at a larger scale. When choosing a county to reside in, people may favor places with a significant presence of their co-ethnics and those that host like-minded residents. Likewise, opportunities to migrate may be turned down if they would lead to settings in which the mover would find themselves socially isolated or targets of discrimination.

Segmented immobility can also arise in more subtle ways: even if individuals do not avoid living with dissimilar others, they may exclude potential migration destinations that are not able to offer the lifestyle and cultural consumption they are used to. Moving from Manhattan to rural Texas, the New Yorker would miss the coffee shop at the street corner, while a Texan migrating in reverse might feel nostalgia for the country music scene back home. Hence, migration between rural and urban areas, and across culturally different states is likely to be disfavored. Racial demographics can also be a determinant of the cultural and economic conditions of a place, where a racially diversified area not only offers a diversity of cultural affordances (as reflected by cuisines and music genres, for example), but also provides vital economic opportunities and ethnic capital for ethnic minorities (Fernández-Kelly 2008; Lee and Zhou 2017; Zhou 1992). Similarly, migrants from rural counties might find themselves excluded from jobs in urban areas because they demand skills hard to obtain in their rural hometown, potentially leading to circulation of poor rural migrants among non-metropolitan counties (Lichter et al. 2022). These together suggest an economic dimension to segmented immobility, in which migration between dissimilar places is suppressed when these places have different economic structures, making it difficult for migrants to utilize human capital accumulated in their place of origin. As services, cultural activities, and modes of production become specialized to a local social ecology, those adapted to both producing and consuming within that ecology will find it increasingly difficult to utilize opportunities in ecologically distinct localities. Together, these mechanisms lead to the following hypotheses:

Hypothesis 1.2: Ceteris paribus, *the more dissimilar counties are in their levels of urbanization, the lower the migration flow between them.*

Hypothesis 1.3: Ceteris paribus, *the more dissimilar counties are in their racial compositions, the lower the migration flow between them.*

The hypothesis of segmented immobility is based on the assumption that most residents and migrants identify themselves with their current residence, which is also the place of departure. However, if we were to suppose that the majority of the migrating population moved to *escape* their current residence in favor of one more to their liking, then migration flows would preferentially occur between dissimilar areas; this would lead to "mobility across segments," in contrast to "segmented immobility." This type of process was proposed by Tiebout (1956) as a mechanism of political sorting, and at the micro-level similar processes have been occur in personnel turnover (Krackhardt and Porter 1986) and cascade-like relocation phenomena (Schelling 1978). We contend that such sorting flows are unlikely to be the major force of the contemporary internal migration in the United States. This is because research has not documented substantive social changes that drove massive redistribution of American population since the fading of the Great Migration of Black Americans in 1970s (Sharkey 2015; Tolnay 2003), and the continuing decline of internal migration for the past decades seems to suggest a scenario of equilibrium, or "an inflection point" (Molloy et al. 2011: 173). Analyses of voting behaviors also reveal that internal migrants tend to hold political orientations consistent with those of their origins (Preuhs 2020). Nevertheless, we consider it as a competing hypothesis to the segmented immobility hypotheses above, and will directly test them in our analysis.

Internal Dynamics: Reciprocity and Perpetuation

The network approach also brings the opportunity to formally examine the interrelationships among migration flows themselves, thereby revealing the internal dynamics of the migration system. This is particularly true for the valued network models used here, which allow us to examine quantitative questions that go beyond the simple presence or absence of migration. Here, we focus on

several mechanisms motivated by prior theory on migration behavior at the micro-level, which lead to hypotheses regarding interdependence among macroscopic migration flows.

We begin by considering the relationship between one migration flow (e.g., from Seattle to Austin) and its opposite flow (e.g., from Austin to Seattle). As has been argued by the transnationalism school in the context of international migration, migration is not a one-way process, but an enduring reciprocal exchange of people, goods, and cultures between sending and receiving countries (Schiller et al. 1995; Waldinger 2013). These same mechanisms could also apply to movement within countries: in his classic work, Ravenstein (1885:187) documented the "universal existence" of "counter-currents of migration" between counties in the United Kingdom, where population not only moved from agricultural areas to commercial and industrial areas, but each of these migration currents corresponded to a current running in the reverse direction. Considering that migration control policies suppress the circulation of international migrants between states (Czaika and de Haas 2017; Massey et al. 2016), we expect even stronger reciprocity of migration flows in the context of internal migration in the U.S., where there is no state control over migration. The reciprocity can arise from the sharing exogenous properties of the bidirectional flow; for example, geographical proximity is a driver of reciprocal population exchange, as it facilitates migration in both directions. Nevertheless, we argue that reciprocity is also an internal dynamic of the migration flow system, such that net of exogenous factors, a larger migration flow in one direction is still associated with a larger migration flow in the opposite direction.

The endogenous, systemic pattern of reciprocity could result from at least two micro-mechanisms in the American migration system. First, migration in one direction actively motivates the flow in the opposite direction. Migrants bring information and social connections from their origin to destination, inspiring and facilitating migration in the opposite direction. Second, return migrants participate in flows in both directions, contributing to the positive association between the pair of flows. For example, Spring et al. (2021) find family ties to be a decisive factor for people separated from their spouses or cohabiting partners to return to their hometowns. von Reichert et al. (2014a,b) show that migrants returning from urban to rural areas are mainly driven by social

connections rather than economic opportunities, and they usually bring people in their family network along when they return. Given the plausibility of both mechanisms, we posit the following macro-level hypothesis:

Hypothesis 2.1: Ceteris paribus, *the flow of migration from county A to county B increases with the flow of migration from county B to county A.*

As is implied above, an important feature underlying the macro-level pattern of reciprocity is the presence of (interpersonal) migrant networks that link persons in the sending and receiving regions, so as literature on transnationalism points out (Lubbers et al. 2020; Mouw et al. 2014; Verdery et al. 2018). Migrant networks, according to the definition of Massey et al. (1993:448), "are sets of interpersonal ties that connect migrants, former migrants, and nonmigrants in origin and destination areas through ties of kinship, friendship, and shared community origin." We have argued that, theoretically, migrant networks should contribute to the reciprocity of migration-flow networks, by migrants bringing resources to destination and triggering population moving in the opposite direction, and by motivating return migrants moving between regions in both directions. Yet, reciprocity is not the only pattern that emerges from migrant networks. As the cumulative causation theory argues, the formation and development of migrant networks are a key contributor to the perpetuation of migration flows, which suggests the presence of inertia (aka. a positive association) of the same migration flow over time (Massey 1990; Massey et al. 1993). Specifically, migrants not only bring information and social connections of origin to their destination, triggering migration in reverse, but also take those kinds of resources from destination back to their origin, by returning home or via communication with nonmigrants back home; this lowers the costs and potentially raises the aspiration of migrating to the same destination, making future migration more likely to happen (Garip 2008; Garip and Asad 2016; Liang et al. 2008; Liang and Chunyu 2013; Lu et al. 2013; Massey et al. 1994a; Palloni et al. 2001). Therefore, we hypothesize the perpetuation of migration flows in the system:

Hypothesis 2.2: Ceteris paribus, *the flow of migration from county A to B increases with the past flow of migration from county A to county B.*

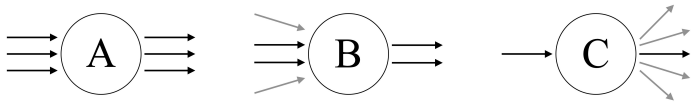


Figure 2. Waypoint Flows
Note: County A, B and C have the same number (six) of associated migration events, but their levels of equality in the in- and out-migration flows vary. This is reflected on their volumes of waypoint flow, three for the most equal County A, two for the medium equal County B, and one for the least equal County C.

Waypoint Flows

We now turn to the internal dynamics at the level of triads, i.e., among three localities (Davis and Leinhardt 1972). Specifically, we examine the *waypoint structure* in the migration flow networks. Similar to a layover airport that mainly serves connecting flights, the "waypoint" is a place where its scales of migrant inflows and outflows are similar to each other. Demonstrated in Figure 2, County A, B, C have the same amount of associated migration events in total (six), but their distributions of immigration and emigration are different. County A is an example of waypoint, where inflows and outflows are evenly distributed, while County C is a counter-example that has few inflows but many outflows, and County B is in between. The difference can be represented by the measure of *waypoint flow*, which is the total amount of migration flows moving in and out of a focal place. When we hold constant the total number of migration events, a high volume of waypoint flow represents a high level of equality between their inflows and outflows. In Figure 2, the volume of waypoint flows for County A, B, C are three, two, one, respectively, indicating that County A has the most balanced inflows and outflows, followed by B and C.

Waypoint flows can arise from chain-like migration processes (Leal 2021), such as stepwise migration and relay migration. Stepwise migration refers to movements of migrants that pass through at least one waypoint before reaching the final destination (Conway 1980). Originally theorized in the classic piece of Ravenstein (1885), stepwise migration has been widely documented to happen under various social contexts (Freier and Holloway 2019; Paul 2011, 2017; Riddell and Harvey 1972), including internal migration in the United States (DeWaard et al. 2016). Stepwise

migration usually happens when the final destination is not directly reachable because of the high financial burden or the hardship in acquiring visas for international migration; migrants respond to this challenge by first migrating to waypoints that facilitate their accumulattion of capital of various kinds before moving to their ultimate stop (Paul 2011). Another migration process that gives rise to waypoints is relay migration, where exodus of local residents leave vacancies in the labor market that attract inflows of migrants (Durand and Massey 2010). Relay migration can also happen in the reverse order, where the influx of migrants triggers outflows of local residents (Leal 2021). The key difference between stepwise migration and relay migration is that the former is about the same migrant taking a multiple-step move, but the latter involves different populations participating in the inflows and outflows of waypoints.[4] The two processes are not distinguishable in aggregate migration flows, but both reflect the interconnectedness of the migration system, where the change of one migration flow could alter another via their shared connection at the waypoint.

While existing literature has studied the migration processes that can generate waypoint flows, less is known about their prevalence in the migration systems. This knowledge gap drives us to further theorize chain-like migration processes by considering them against other migration processes. Since migration is an arduous undertaking with substantial risks, costs and barriers (Carling and Schewel 2018; Liang et al. 2008; Schewel 2020), prolonging one-step migration into stepwise is not a desirable choice unless necessary. Compared to international migration, internal migration in the U.S. is usually more affordable and less constrained by state regulations; an American internal migrant is thus less likely to opt for stepwise migration than a Filipino who wishes to settle in Spain. Relay migration is not a universal pattern, either. It requires substantial inflows or outflows that can alter the local labor and housing market or socio-political contexts to trigger further migration flows. This means that waypoint flows arising from relay migration is conditioned on uncommon incidents such as major economic shocks or environmental disasters that bring mass population movements.

Moreover, a *deficit* in waypoint flows can also be a structural signature of inequality in migration flow networks, where the majority of counties either receive many migrants but send few, or

send many migrants but receive few. This imbalance between in- and out-migration flows can arise when the difference in the level of attractiveness across places remain unaccounted for; in this case, a county is either popular so to attract and retain migrants, or the reverse. A lack of waypoint flows can also occur endogenously. For instance, potential migrants may take current levels of migration rate as social or economic signals about the long-term desirability of an area, and adjust their own decisions accordingly. This tendency creates a feedback loop in which influx of migrants to an area leads potential out-migrants from the area to instead remain, which in turn feeds an imbalance between in- and out-migration ($in > out$) that motivates yet more potential migrants to move in. By turns, this same mechanism may lead to a Schelling-like exit cascade (Schelling 1978), in which an initial out-migration shock both encourages further exit from those now in the location and makes the location appear less-desirable to potential in-migrants, thus leading to poorer in/out balance ($in < out$), and further net out-migration.

Clearly, then, there are interesting and plausible hypotheses in both directions. For simplicity, we hypothesize a high-waypoint scenario, reflected by a balanced distribution of inflows and outflows:

Hypothesis 3: Ceteris paribus*, the inflows of migration to a county increase with its outflows.*

It should be noted that the waypoint flow is a network structure related to but distinct from the transitive hierarchy studied in some international migration network research (Leal 2021; Windzio 2018). Both are triadic structures concerning migration flow among three places (i, j, k). The waypoint flow is the backbone of the transitive hierarchy, as the former considers migration flows of $i \to j$ and $j \to k$, while the latter involves the co-presence of $i \to k$ flow. This means that networks with a lack of waypoint flow will have few closed transitive triads ($i \to j, j \to k, i \to k$).[5] We thus focus on the more fundamental waypoint flow structure to explore the more basic form of endogenous mechanism in the migration network.

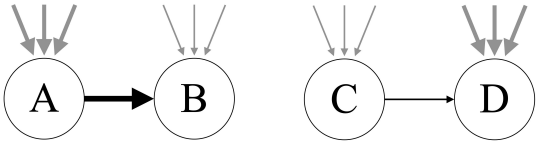


Figure 3. Hypothesized relation between internal and international migration
Note: vertical grey arrows denote international immigration flows and horizontal dark arrows denote internal migration flows. Arrow width denotes the magnitude of migration flows. According to the hypothesis of Frey (1995a), larger population are expected to migrate from County A that has high immigrant inflows towards County B that has low immigrant inflows, while less population would leave County C that has low immigrant inflows towards County D that has high immigrant inflows, net of other factors.

Internal Migratory Response to Immigration

Lastly, this paper considers the relationship between international migrant (i.e., immigrant) inflows and internal migrant flows in the United States. Debates about the impact of immigration on internal migration provoked much research in 1990s, which provided insights about the demographic and economic influence of immigration, the structure of labor markets, and the social cohesion of American society. Frey (1995a) hypothesized that immigration to the U.S. would lead to demographic balkanization, in which immigrant inflows trigger outflows of internal migrants and deter their inflows. Figure 3 visualizes this hypothesis from the perspective of internal migration flows, where larger population are expected to migrate from County A that has high immigrant inflows towards County B that has low immigrant inflows, while less population would leave County C that has low immigrant inflows towards County D that has high immigrant inflows, net of other factors. This mechanism was proposed to lead to a "balkanized" regionalization of the U.S., with immigrants and natives increasingly segregated in different regions.[6] Empirical findings were inconclusive about the relationship between internal and international migration flows, with some supporting evidence for Frey's (1995a) hypothesis (Borjas 2006; Frey 1995a,b; White and Liang 1998), and other opposing evidence (Card 2001; Kritz and Gurak 2001; Wright et al. 1997). This paper revisits this debate with new data about migration in 2010s of all U.S. counties. Following Frey's (1995a) proposal, we hypothesize, from the perspective of internal migration flows, that:

Hypothesis 4: Ceteris paribus, *an internal migration flow increases with international immigration inflow in the sending county, but decreases with international immigration inflow in the receiving county.*

DATA AND METHODS

Valued TERGMs

We use the valued temporal exponential-family random graph models (valued TERGMs) to study the intercounty migration-flow network within the United States. Exponential-family random graph models (ERGM) offer a flexible framework that describes the probability of observing certain network structure as a function of their nodes' covariates, edges' covariates, and the dependence structure among edges (Hunter et al. 2008; Wasserman and Pattison 1996). This empowers us to simultaneously model the characteristics of areal units (nodes' covariates), the relational linkages (edges' covariates), and the internal dynamics (dependence structure) hypothesized to characterize migration-flow networks. Previous research has employed ERGMs in a wide range of social network settings, including friendship networks in schools (Goodreau et al. 2009; McFarland et al. 2014; McMillan 2019), inmate power relationships in prison (Kreager et al. 2017), collaboration networks in firms (Srivastava and Banaji 2011), online social networks (Lewis 2013, 2016; Wimmer and Lewis 2010), and various types of gang networks (Lewis and Papachristos 2019; Papachristos et al. 2013; Smith and Papachristos 2016). While most studies model social relations as binary networks (i.e., encoding only whether or not relationships exist), it is more accurate and informative to model migration-flow systems as valued networks, where edges represent the size of population migrating between county pairs. Although valued ERGMs (VERGMs) are to date less well-studied than binary ERGMs, we employ the count-data ERGM framework of Krivitsky (2012) to capture migration rates in a quantitative fashion. Our model also incorporates temporal effects (the perpetuation pattern), making it a valued temporal ERGM, or valued TERGM).

We detail the model setup, computation methods and procedures in Part B of the supplement.

We also develop and report a model adequacy check for VTERGMs, detailed in Part D of the supplement.

Knockout Experiments

Exploiting our ability to quantitatively model the magnitude of migration flows using VTERGMs, we perform *in silico* "knockout experiments" to show the impact of modelled social mechanisms in influencing the size of the migrant population, tackling the question of how particular social forces give rise to immobility. Originating and widely used in the experimental sciences (Hall et al. 2009; Vogel 2007), this way of thinking has also been applied in the social sciences (e.g., Han et al. 2021; Lakon et al. 2015; Xie and Zhang 2019), especially in the context of agent-based modelling (Miller and Page 2009). For social science research, the knockout experiment can be considered as a model-based thought experiment (*Gedankenexperiment*, Einstein et al. 1935), where we use models to predict social outcomes of interest (e.g., total number of migrants) under a counterfactual scenario where certain social mechanisms are removed (e.g., the political segmentation effect) while other factors are held constant. This approach is particularly powerful for nonlinear, systemic models like those used here, where seemingly small, local effects can have global consequences.

Our knockout experiments are performed as follows. Starting with a VTERGM calibrated using empirical migration data, we compute the total expected number of intercounty migrants when either the political segmentation mechanism (*per se*) or all of the three segmentation mechanisms (jointly) are knocked out, and compare this number with the observed migrant population size. The differences in total migrant population between these scenarios thus offer an insight about the scale of mobility suppression from these segmentation mechanisms - i.e., if we could "turn them off," what would we hypothetically expect to see? The counterfactual scenario was simulated by the Markov chain Monte Carlo (MCMC) algorithm based on the network model with zero coefficient values for the specified knockout social effects.[7] Since the network model specifies the dependence structure between migration flows, it accounts for both direct impacts of the segmentation between each county pair on their own migration flows, and the indirect impact arising from

the internal dynamics of migration systems that spillover this exogenous impact. It thus offers a systemic depiction of the segmented immobility pattern.

Data

We analyze the inter-county migration flow data for from the American Community Survey (ACS). As a political unit with reliable demographic and economic data, counties serve as a level of geographical area that effectively describes the social contexts of residents such as political environments and rurality (Lobao and Kelly 2019; Mueller and Gasteyer 2023; Schroeder and Pacas 2021). Movement across a county boundary is a frequently-used definition of internal migration in the literature (Brown and Bean 2016; DeWaard et al. 2020; Hauer 2017; Partridge et al. 2012). Administered by U.S. Census Bureau, ACS surveys respondents' location of residence one year ago and estimates the population size that migrated between each pair of counties each year.[8] Their released data reports the averaged annual migrant counts in a five-year time window in order to have enough monthly samples for reliable estimation at the inter-country level. The outcome of interest is the count of migrant population flowing between 3,142 counties in the United States during 2011-2015.

The explanatory variables are from 2010 United States Census and ACS 2006-2010. Specifically, the intercounty distance was calculated based on the 2010 Census by National Bureau of Economic Research (2016). We use presidential election turnout in 2008 to indicate the political climate of each county (MIT Election Data and Science Lab 2018). Data sources for each covariate are listed in Part A of the supplement.

Variables

Dependent edge variable. The model predicts the count of migrants moving between each directed pair of counties during 2011-2015 from the American Community Survey. Because the count-valued ERGM effectively operates through a logarithmic link (see Krivitsky 2012), we are able to directly predict untransformed migrant counts in the model.

Dissimilarity score for segmented immobility. The segmented immobility thesis contends that less migration happens between places with different political climates, levels of urbanization, and racial compositions. To test the hypotheses, we measure the dissimilarity within each pair of counties along these dimensions as edge covariates for migration flows.[9] For difference in political climates, we follow Liu et al. (2019) and calculate the absolute difference in percentage of votes for the Democratic candidate in the 2008 presidential election, a behavioral measure of partisanship.[10] For levels of urbanization, we calculate the absolute difference in percentage of population residing in rural areas, a standard urbanization measurement reported in 2010 Census. For racial/ethnic composition, we use a function of the sum of absolute differences in population share for each racial category. Formally, we describe relationship between counties A and B by

$$R_{AB} = \frac{1}{2}\sum_{i=1}^{n}\left|\frac{P(A)_i}{P(A)} - \frac{P(B)_i}{P(B)}\right|$$

where R_{AB} is the dissimilarity score of racial composition between county A and county B, $P(A)$ is the total population size of county A and $P(A)_i$ is the population size of the i-th racial group in county A. We follow the Census to consider the following five racial/ethnic categories, Hispanic or Latino, Non-Hispanic Black or African American, Non-Hispanic Asian, Non-Hispanic White, and population with the other racial identifications. The difference is divided by two to make the theoretical value of the score range from 0 to 1. The higher the dissimilarity score, the more different the two counties are in the measured dimension, and the less migration is expected according to the hypotheses.

Network covariates. We utilize the mutuality statistic in the ergm.count R package to measure reciprocity in migration flows (Krivitsky and Butts 2013).[11] A positive coefficient for indicates reciprocity within the network, such that a large migration flow is more likely to have a larger counter current rather than a smaller one, *ceteris paribus*.

The model also includes the number of migrants in the past 5-year window during 2006-2010 in log scale from ACS as an edgewise covariate, to account for the association of migration flows

over time, utilizing the temporal feature of TERGMs. A positive coefficient for this term suggests the perpetuation of migration flows over time, while a negative coefficient suggests negative dependence between past and present flows.

Waypoint flow is captured by the summation of the volumetric flow for each county in the network. Intuitively similar to the notion of the flow volume "through" or "across" an areal unit in the field of fluid mechanics, the flow associated with a given unit is the minimum of its total inflows and its total outflows.[12] A positive coefficient for the flow term indicates that the observed network has larger volumes of waypoint flows than would be expected given all other mechanisms and covariates specified in the model, suggesting a relatively equal distribution of in- and out-migration flows across counties, and a negative coefficient would indicate otherwise.

To examine the relationship between internal and international migration flows, for each inter-county migration flow, the model measures its associations with the total immigrant inflows of its sending and receiving counties in the same time window (2011-2015). The international immigrant population is transformed by taking the natural logarithm.

Demographic covariates. The model also accounts for areal characteristics that might influence intercounty migration. These include demographic characteristics of the sending and receiving counties, from basic geo-demographic statistics to demographic compositions.

Classic models from spatial econometrics (a.k.a. the gravity model) suggest that migration rates are positively associated with the population sizes of the sending and receiving regions, but negatively associated with their distance, with a general power law form (Boyle et al. 2014; Poot et al. 2016; Zipf 1946, 1949). Such models can be expressed by a linear combination of population and distance in the log space. Formally,

$$log(M_{AB}) = \beta_0 + \beta_1 log(P_A) + \beta_2 log(P_B) + \beta_3 log(D_{AB}) + \varepsilon$$

where M_{AB} is the migration volume from A to B, P is the regional population, D is the inter-regional distance, β is a covariate vector, and ε is the residual. Almquist and Butts (2015) suggest that this

may arise from the volume of interpersonal contacts between regions, which also frequently scales in power law form. Although we do not use a regression model of this type here, we emulate this class of effects within our own model by incorporating (1) the log populations for the sending and receiving counties and (2) the log distance between counties (in kilometers) as predictors of inter-county migration rates; this means that our models can be considered as an extension of the gravity model. We also include population densities of sending and receiving counties (in thousand people per squared-kilometer), since Cohen et al. (2008) has shown that population density is a critical factor in predicting international migration flows. We use data from the 2010 Census for the covariates listed above.

For demographic composition, the model first considers the age structure of sending and receiving counties, as Kim and Cohen (2010) found that migrants are more likely to leave younger countries towards older countries in the context of international migration. Using the 2010 Census, the potential support ratio (PSR) equals to the ratio of population aging 15-64 over population aging 65+, which is the inverse of dependency ratio in demography literature; the higher PSR, the younger the population.

Racial composition could influence the mobility of population as well, as extant literature found different patterns of internal migration between racial groups (Crowder et al. 2012; Sharkey 2015). Hence, besides the dissimilarity of racial composition between counties, we also consider the racial composition of the sending county to account for the varying mobility of different groups, as measured by the proportion of each racial category in the population.

Economic covariates. Economic structures of origins and destinations could potentially influence their migration flows. Since renters on average are more mobile than house owners (Frey 2009; Molloy et al. 2011) even after controlling for demographic and socioeconomic factors (Jia et al. 2022), the model includes the percentages of housing units occupied by renters for both origin and destination, using 2010 Census data. The model also controls the percentage of population with a college degree using the 2006-2010 ACS. This is because human capital may offer greater ability and opportunities for migration, and previous analysis found that population with higher

education attainments have higher migration rates in the U.S. (Frey 2009).

Neoclassical economic theory predicts that people migrate towards economic opportunities (Massey et al. 1993; Todaro 1976). The theory also predicts that regions with more economic opportunities will send more migrants, since their population have more capital to finance their migration (Massey and Espinosa 1997). We thus include the unemployment rate of the origin, and the difference in the unemployment rate between the destination and the origin. In combination of neoclassical economic theory and the aspiration-ability model (Carling 2002), we hypothesize that more migration will come from counties with lower unemployment rate given their greater ability to move, and more migration will happen when the destination has lower unemployment rate than the origin, offering more economic opportunities and higher aspiration for migration. Similarly, the models incorporate the logarithm of median monthly housing costs of the origin and the difference in log housing costs between destination and origin.

Geographical covariates. Besides distance between counties, the model also controls for regional differences in mobility. Previous research found that migration rates and their trends in different parts of America vary significantly (Frey 2009). We believe that the regional difference may not be fully explained by difference in social contexts indicated by the covariates above. Dummy variables are created to indicate whether the origin and destination is in the West, the Midwest, the South, with the Northeast as the reference group, based on the definition of U.S. Census Bureau (2013).

Administrative boundaries are likely to influence migration flows as well. Charyyev and Gunes (2019) found that, marginally speaking, the majority of inter-county migration in the U.S. happens within a state, and in this paper we further examine whether state boundary influences migration flows after controlling for distance and dissimilarity between counties. Intrastate intercounty migration could be more prominent than cross-state migration because compared to intrastate migration, the cross-state migration creates extra burdens ranging from adaptation to unfamiliar legal and cultural environments, to navigation of administrative procedures such as change in occupational licensing for workers in certain occupations (Johnson and Kleiner 2020). Yet, the opposite

hypothesis is plausible under the consumer-voter model, which contends that people vote by their feet (Dye 1990; Tiebout 1956); as means of pursuing favorable policies, cross-state migration is more effective if people migrate to seek lower tax rates or more welcoming policies and climates for immigrants (Preuhs 1999; Schildkraut et al. 2019). The model creates a dummy variable indicating whether the two counties are affiliated with the same state. A positive coefficient would suggests that intrastate intercounty migration is more prominent, and a negative coefficient suggests that inter-state migration is more prominent.

Variable Setup. We report two models in the following results section. The first model contains every covariate except the rural dissimilarity score, which is later included in the second model, the full model. Since the level of urbanization is strongly associated with their political environment, comparison between the two models could reveal how much of the total effect of political dissimilarity might be explained by their difference in the level of urbanization. Besides the sum term serving as an intercept, we add to models a term that counts the number of nonzero dyads of the network to account for the zero-inflation of migration flow data (Krivitsky and Butts 2013). Its negative coefficients in Table 1 indicate the sparsity of migration flow network, that a county pair is more likely to have no migrants moving between than otherwise, even after controlling for all the covariates in the model. Summaries of descriptive statistics and data sources are attached in Part A of the supplement.

RESULTS

Bivariate Analyses of Migration and Political Division

To explore the pattern of segmented immobility by political orientation, we first perform bivariate analyses between intercounty migration and political division, as visualized in Figure 4. We divide counties into two broad groups, Democratic counties and Republican counties. Democratic counties are counties where the Democratic candidate (Obama) received more votes than the Re-

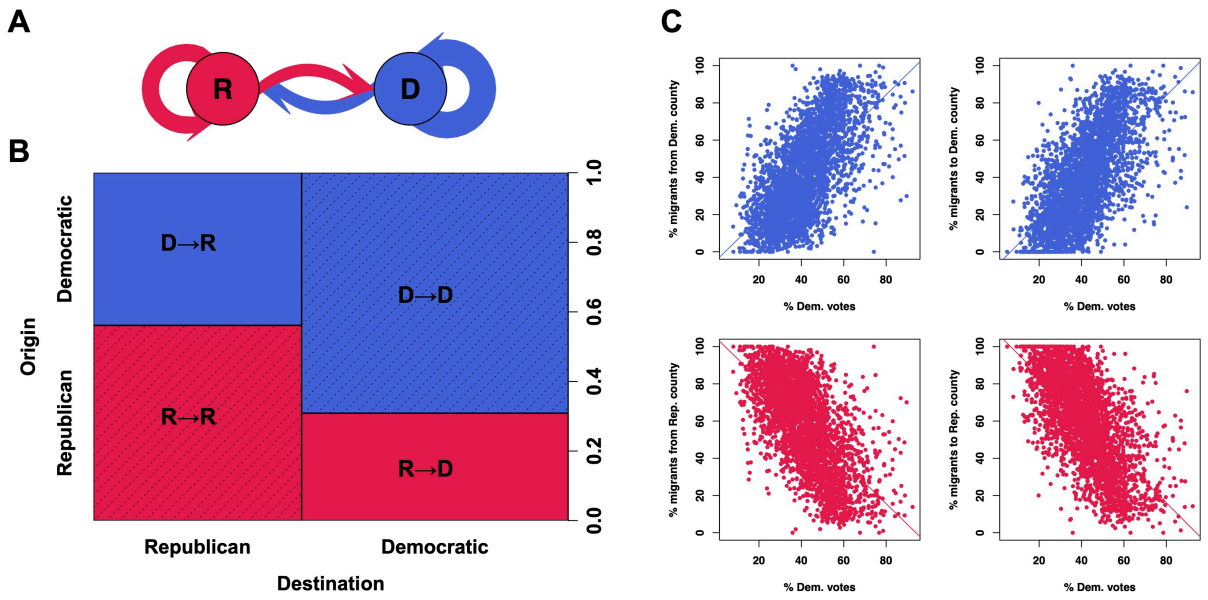


Figure 4. Immobility from political division
Note: The sociogram (A) represents the magnitude of migration flow within and between Democratic counties (node D in blue) and Republican counties (node R in red), which is proportional to the width of the edge. The spineplot (B) represents the magnitude of migration flow within and between the two groups by the area of each block. The shaded blocks represent migration within each group. Scatterplots (C) show the relationship between percentage of Democratic votes in 2008 of a county and the composition of its in-migrants and out-migrants. The lines are fitted bivariate linear regression lines.

publican candidate (McCain) in the 2008 presidential election, and vice versa for the Republican counties. The sociogram in Panel A of Figure 4 visualizes the magnitude of migration within and between Democratic and Republican counties, which is proportional to the width of edges. The sociogram shows that migration flows within each group has thicker edges than flows between, suggesting that more migration happens from one Democratic county to another, or from one Republican county to another, than between a Democratic county and a Republican county. The spineplot in Panel B represents the magnitude of migration flow within and between groups by the area of each block. The shaded blocks are migration happening within Democratic or Republican county groups, suggesting again that more migration happens on either side of the party line than across it. The color of each block indicates whether the origin of the migration flow is from a

Democratic (blue) county or a Republican (red) county. The spineplot indicates that only 31% of the migrants moving into a Democratic county come from a Republican county, and just 44% of the migrants moving into a Republican county come from a Democratic county.

Panel C of Figure 4 visualizes the relationship between the percentage of the Democratic votes in the 2008 election and the composition of the in-migrants and out-migrants for each county. The upper left panel shows that the higher the Democratic vote in 2008, the larger the proportion of migrants coming from a Democratic county, and the smaller the proportion of migrants coming from a Republican county, as shown in the lower-left panel. Similarly, the right-hand column suggests that a larger share of 2008 Democratic votes within a county is associated with a larger proportion of out-migrants moving to a Democratic county, and a smaller proportion to a Republican county. Overall, the figures reveal a clear and strong pattern of political sorting, where less population migrate between counties with distinct political environments than those with similar political environments.

Segmented Immobility

The bivariate analysis is suggestive that intercounty migration is immobilized by political divisions in the United States. We further examine this using VTERGMs that incorporate the demographic, economic, geographical and political factors at the county and inter-county levels, together with explicit specifications of internal dynamics of migation systems. Table 1 displays the results. Model 1 suggests that, holding all other factors constant, a larger difference in political environments between counties predicts less migration between them. Since the political environment is associated with the level of urbanization of a county (Cramer 2016), Model 2 further includes the dissimilarity of urbanization between counties. From Model 1 to Model 2, the effect size of political dissimilarity becomes modestly smaller, suggesting that the effect of political difference can be partly (but not completely) explained by their difference in the level of urbanization. The smaller BIC of Model 2 further indicates that difference in the level of urbanization is effectively explaining the variation in the magnitude of migration flows. Nonetheless, in Model 2, larger political dissimilar-

ity is still a statistically significant predictor of less migration between counties, offering empirical evidence for Hypothesis 1.1. Holding other factors constant, a pair of counties with 10% larger difference in 2008 voting outcome is expected to have 2.5% (i.e., $[1-\exp(-0.256\times 10\%)]$) fewer migrants than another county pair. Similar to political segmentation, Model 2 also reveals that larger differences in levels of urbanization and racial compositions of two counties predict fewer migrants moving between, holding other factors constant, lending support for Hypotheses 1.2 and 1.3. The VTERGM results do suggest that migration is inhibited between places with dissimilar political contexts, levels of urbanization, and racial compositions.

Table 1. Valued TERGMs for Inter-county Migration Flows, 2011-2015

	Model 1		Model 2	
	Estimate	SE	Estimate	SE
Segmented Immobility				
Political dissimilarity	-.368***	.007	-.256***	.007
Rural dissimilarity			-.399***	.004
Racial dissimilarity	-.361***	.006	-.217***	.006
Network Patterns				
Mutuality	.054***	.002	.045***	.002
Log(past migrant flow)	.303***	<.001	.300***	<.001
Waypoint flow	-.014***	.001	-.015***	.001
Destin.log(immigrant inflow)	.062***	.001	.056***	.001
Origin.log(immigrant inflow)	.040***	.001	.035***	.001
Demographics				
Destin.log(population size)	.351***	.002	.351***	.002
Origin.log(population size)	.370***	.002	.373***	.002
Destin.log(population density)	-.077***	.001	-.083***	.001
Origin.log(population density)	-.062***	.001	-.069***	.001
Destin.PSR	.018***	.001	.017***	.001
Origin.PSR	.013***	.001	.013***	.001
Origin.P(White)	(reference group)			
Origin.P(Hispanic)	-.012	.007	-.064***	.007
Origin.P(Black)	.147***	.008	.117***	.008
Origin.P(Asian)	.408***	.020	.467***	.020
Origin.P(other race)	1.031***	.015	.993***	.015
Economics				
Destin.P(renter)	.405***	.011	.348***	.011
Origin.P(renter)	.507***	.012	.476***	.012
Destin.P(higher education)	.327***	.011	.359***	.011

Table 1. *(continued)* Valued TERGMs for Inter-county Migration Flows, 2011-2015

	Model 1		Model 2	
	Estimate	SE	Estimate	SE
Origin.P(higher education)	.157***	.012	.153***	.012
Difference.log(housing costs)	-.135***	.004	-.153***	.004
Origin.log(housing costs)	-.248***	.005	-.277***	.005
Difference.P(unemployment)	-1.305***	.040	-1.300***	.040
Origin.P(unemployment)	-3.039***	.052	-3.012***	.052
Geographics				
Log(distance)	-.563***	.001	-.568***	.001
Same state	.501***	.002	.510***	.002
Northeast	(reference group)			
Destin.South	.258***	.003	.253***	.003
Origin.South	.047***	.003	.046***	.003
Destin.West	.384***	.004	.374***	.004
Origin.West	.193***	.004	.184***	.004
Destin.Midwest	.203***	.003	.197***	.003
Origin.Midwest	.085***	.003	.080***	.003
Baseline				
Sum	-1.609***	.040	-1.193***	.040
Nonzero	-13.966***	.028	-13.917***	.028
BIC	2,221,363		2,210,125	

Note: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$ (two-tailed tests).

Table 2. Migrant Population Sizes under Observed and Knockout Scenarios

	Total Migrants	Increment in Count and Rate	
Observed	17,176,675		
Remove political segregation	17,965,336	788,661	4.6%
Remove all segmentation	21,741,021	4,564,346	26.6%

To quantify the contribution of segmented effects to immobility, we perform knockout experiments to compute the total migrant population under counterfactual scenarios where these effects are inoperative, and compare that with the observed scenario. Table 2 shows that when the political segregation effects on migration flows were knocked out, the expected intercounty migrant population each year would increase by 789 thousand, 4.6% higher than the observed. At the absence of all three segmentation patterns, we would expect to observe 26.6% more internal migrants in the United States, that is 4.56 million more people moving from one county to another each year.[13]

Results of the VTERGMs and knockout experiments together suggest that segmented immobility serves as a critical and substantial social mechanism behind the immobility of the contemporary American society. These social mechanisms may be partly driven by economic forces (although supplementary analysis shows that dual labor and housing markets make little impact on the described segmentation pattern, see Part C of the supplement); it may also reflect people's preference for residing in an environment that is culturally and politically familiar to them. This tendency not only implies social cleavages along party lines, between urban and rural lands, and across communities with varying racial demographics; it could also contribute to a growing geographical segmentation along those lines. As has been known since the classic works of Sakoda and Schelling, even a small preference for homophily can lead to substantial segregation in residential settlement patterns (Fossett 2006; Sakoda 1971; Schelling 1969).

Network Dynamics Influencing Migration Flows

The VTERGMs also consider the network patterns of the migration flow system. That all coefficients are significant in the *Network Patterns* section in Model 2 of Table 1 confirms that they play a significant role in determining the directions and magnitudes of intercounty migration flow. In Model 2, the positively significant mutuality term confirms Hypothesis 2.1, that reciprocity is present in the migration-flow networks: a larger flow from county A to B is positively associated with a larger flow from county B to A, holding other effects constant. Joining research on global migration and intercounty migration in U.K. (Ravenstein 1885; Windzio 2018), we show that reciprocity is also a network pattern found within U.S. migration. It is interesting to note that some prior studies not observe reciprocity effects in their analyses (Desmarais and Cranmer 2012; Windzio et al. 2019); this might come from omission of some regional characteristics that influence the attractiveness of regions to migrants, or their operation of data transformation for the migrant count variable. Future research may replicate the analysis of reciprocity using count-data network models under various social contexts to understand whether reciprocity is a prevalent phenomenon, or can be suppressed by some social forces.

Model 2 also reveals that a larger migration flow during 2006-2010 is significantly associated with a larger migration flow during 2011-2015, even after holding all exogenous and endogenous factors constant. This confirms Hypothesis 2.2 regarding the perpetuation of the migration flow system, showing that migration-facilitating mechanisms offer the system its own momentum, promoting future migration net of exogenous factors such as demographic structures of a region (de Haas 2010).

The significantly negative coefficient of the flow term indicates a lack of waypoint structures of inter-county migration, refuting Hypothesis 3. The negative waypoint flow effect implies that relatively little migration is proceeding in the chain-like manner such as stepwise and relay migration. After holding other factors constant, counties generally have an imbalance or inequality in the scales of their migration inflows and outflows, either sending many migrants but receiving few, or receiving many migrants but sending few. This may represent emergent attractiveness effects, in which in-migration makes a county seem more attractive to other possible migrants, and out-migration makes a county seem correspondingly less attractive. It may also reflects unobserved heterogeneity in attractiveness arising from other factors; the specification of waypoint flows in the model thus controls for this possible source of autocorrelation, beyond its substantive interest.

Note that the inequality identified by a lack of waypoint flows in this inter-county migration network is different from the inequality captured by an abundance of transitive hierarchy in other cross-national migration networks (e.g., Leal 2021). The transitive hierarchy requires many waypoints serving as the "mildly structurally attractive position," between the highly and the minimally "structurally attractive positions" (Leal 2021: 1086). In analogy, that implies a multi-layer hierarchy of the global system with countries positioned in the core, the semi-periphery and the periphery (Wallerstein 2011). On the contrary, in this network with a lack of waypoint flows, there is an *absence* of semi-periphery areas serving as waypoints between the core and the periphery; in comparison with the international migration system, the U.S. migration system is relatively bipolar, with counties tending to be, *ceteris paribus*, either structurally attractive or unattractive, with few in the middle ground.

The model also examines the relationship between internal and international migration. It shows that larger immigrant inflows from 2011 to 2015 are positively associated with larger intercounty inflows and outflows in the same period. This finding does not correspond to either side in the debate about internal migratory response to immigration, which contends that large immigrant inflows are either associated with small internal migrant inflow and large outflows, or not associated with internal migrant flows. Rather, the results suggest that counties with large immigrant inflows are active in *both* sending and receiving intercounty migrants. Further, the larger coefficient of destination effect than the origin effect suggests that, increasing immigrant inflows to a county is associated with larger increase of internal inflow than internal outflow. In other words, immigration is actually associated with net population increase from internal migration. Overall, the finding shows a common mobility pattern for internal and international migration, wherein counties popular among international immigrants are also popular in receiving and active in sending internal migrants.[14]

Demographic, Economic, and Geographic Determinants of Migration

Alongside segmented immobility and network patterns, the models also consider other factors that could influence intercounty migration. For demographic characteristics, Model 2 confirms findings from spatial econometrics (gravity) models that population sizes in both sending and receiving regions are positively associated with migrant flow (Boyle et al. 2014; Zipf 1946, 1949). A 10% increase in destination's population size is associated with a 3.4% (i.e., $[1.1^{0.351} - 1]$) increase in the number of migrants, and a 10% increase in origin's population size is associated with a 3.6% (i.e., $[1.1^{0.373} - 1]$) increase in the number of migrants, holding other factors constant. Population density has a significantly negative effect for both the number of in-migrants and out-migrants, holding population size and other factors constant. One possible mechanism is that higher population density leads to larger shares of local connections for their residents (Butts et al. 2012; Hipp et al. 2013; Thomas et al. 2022), where more job transitions and housing transactions can happen locally thanks to these connections, reducing migration across county borders.

With respect to demographic composition, larger migration flows are significantly more likely to be observed between counties with younger populations, in line with the migration schedule literature finding that younger adults are more mobile than older adults (Raymer and Rogers 2007; Rogers and Castro 1981). The model also shows that counties with larger shares of Hispanic population tend to send fewer migrants, but counties with larger shares of Non-Hispanic Black, Non-Hispanic Asian and Other races populations tend to send more intercounty migrants. Note that these effects do not directly describe the mobility of each racial/ethnic population, since they are predicting the magnitude of migration flow for all racial and ethnic populations. Decomposing migration flows into migrants of each racial/ethnic population is necessary to further reveal the variation of mobility between people with different racial/ethnic identities.

Economic covariates in Model 2 show that larger migration flows exist between counties with higher shares of renters and people with college degrees, consistent with previous literature observing that renters and people with higher education credentials are more mobile than their counterparts (Frey 2009). We also see that larger migration flows happen when the route offers greater declines in housing costs, indicating a tendency of mobility towards cheaper housing (Plantinga et al. 2013). Holding other factors constant, counties with lower housing costs have higher outmigration. This might be due to the better financial conditions renters have in low housing cost areas, enabling them to move and relocate. It is also compatible with previous findings that lower housing equity is associated with higher mobility rates (Coulson and Grieco 2013). For unemployment rates, the model suggests that the lower the unemployment rate at the origin, and the larger the decline in unemployment rate from origin to destination, the more intercounty migration. These results are compatible with the cost-benefit model of the neoclassical economic theory of migration that population move towards economic opportunities (Todaro 1976), and that more economic opportunities financing migration makes migration more likely to happen (Massey and Espinosa 1997). The relational approach employed here enables empirical analysis of the aspiration-ability model (Carling 2002; Carling and Schewel 2018), revealing that both the aspiration, as influenced by the relative economic conditions of origin and destination, and the ability, as influenced by the

economic conditions of the origin, matter to migration behaviors.

In terms of geographical factors, the model suggests a negative association between distance and number of migrants flowing between two counties, as the gravity model predicts (Zipf 1946, 1949). A 10% increase in distance between two counties is associated with a 5.3% (i.e., $[1 - 1.1^{-0.568}]$) decrease in intercounty migration. Administrative boundaries also influence migration flows; migration flows within the same state are expected to be larger than those across states, holding other factors constant. Additionally, different U.S. regions have varying mobilities. The model indicates that compared to the Northeast, every other region receives and sends more intercounty migrants, *ceteris paribus*. This suggests the existence of some latent characteristics inhibiting the mobility of the Northeast, which deserves more examination in future work.

Lastly, to check the model adequacy, we simulate networks based on Model 2 (the full model) in Table 1 using MCMC algorithms. We then calculate the total in-migrant and out-migrant count for each county, and compare the observed distribution with the simulated distribution. We find that the fitted model recapitulates the county-level migration data (see Part D of the supplement). We also calculate the Pearson's correlation between observed and simulated distributions, which are all above 0.95. We conclude that the model effectively reproduces the quantitative features of observed migration flow networks.

DISCUSSION AND CONCLUSION

This paper offers a comprehensive analysis of the inter-county migration structure encompassing not only economic, demographic and geographical factors, but also political, cultural factors and internal dynamics of the migration system. Network models reveal a pattern of segmented immobility in America, in which less migration happens between counties with dissimilar political environments, levels of urbanization, and ethnic/racial compositions. Yet, we do not observe segmentation between internal migrants and international immigrants; rather, the model shows that counties active in receiving many international immigrants are active in both sending and receiving

many internal migrants as well. Our analysis also suggests the significance of internal dynamics of the migration flow system; we observe strong patterns of reciprocity and perpetuation, along with a suppression of waypoint structure. These results lend empirical evidence to the systemic theory of migration (Bakewell 2014; de Haas 2010; Mabogunje 1970; Fawcett 1989), showing that the population flows assemble an interdependent network system that carries its own momentum.

This paper identifies segmentation as a critical mechanism behind population immobility in the contemporary American society, which could potentially have deterred millions of people from migrating each year, as suggested by the knockout experiments. This finding implies people's tendency of choosing residency in localities that match with their political affiliations and socio-cultural attributes, potentially leading to geographical segmentation between people with different political identities (Brown and Enos 2021) and increasing the homogeneity of their social relations (DiPrete et al. 2011). Such sorting could possibly reinforce political polarization (DellaPosta and Macy 2015), and can also serve as a mechanism that maintains and even exacerbates residential segregation along other dimensions (Fossett 2006; Sakoda 1971; Schelling 1969). While classic analyses of segregation have focused on local communities within urban areas (Bishop and Cushing 2009), the effects seen here could potentially contribute to macro-level segmentation across the whole country (Liu et al. 2019). From a migration perspective, although internal migration in the U.S. does not involve border-crossing in international migration or other forms of governmental restrictions (such as the household registration system in China, *hukou*), population movement is never free of constraints. Rather, as our analysis shows, Americans today are separated by the invisible borders and walls standing along the party lines, at the midway between rural and urban landscapes, and over the gap across communities with varying racial demographics.

The analytical framework in this paper provides an example of structural and systemic analysis of mobility and immobility, broadly defined. The relational approach connects the perspectives of emigration and immigration to examine how characteristics of origin and destination *jointly* influence migration, which enables revelation of the segmented immobility in the U.S. migration system. The formal specification of the interdependence between migration flows under the ERGM

framework identifies the structural signature of networks, reflecting the internal dynamics of migration systems. The knockout experiment offers model-based insights into how the system might react to social change. Lastly, leveraging advances in scalable VERGM estimation and simulation allows quantitative analysis of the magnitude of population flows and their determinants in large social systems. The applicability of this framework extends beyond the population movement between geographical areas, encompassing mobility in the occupational system for the study of social stratification and mobility (Cheng and Park 2020), the exchange of personnel between organizations (Sparrowe and Liden 1997), and the migration of scholars between institutions and research domains in the sociology of knowledge (Burris 2004; Gondal 2018; McMahan and McFarland 2021).

While our study enables a much richer examination of the mechanisms driving or inhibiting internal migration at a larger scale than what has been possible in extant literature, it is not without its own limitations. First, as a macrosociological study about the "functioning of a social system" (Coleman 1986: 1312), this paper informs an aggregate-level social phenomenon, i.e., population immobility. While analysis of the migration flow network facilitates a systemic understanding of migration and its relation to segmentation from a holistic viewpoint, it does not directly describe the patterns of individual migration behavior. Although we can test for the structural signatures of such micro-level processes, unpacking those fine details requires information on decision making and behavior patterns at the individual level. For example, distinguishing stepwise migration and relay migration requires data about the migration trajectories of individual migrants. Studies like this are hence complementary to micro-level analyses (both quantitative and qualitative) that could shed further light on processes at the individual and household levels (e.g., DeLuca et al. 2019; Fitchen 1994; Lichter et al. 2022; Quillian 2015). Research that aims to bridge individual behaviors and aggregate social outcomes are deemed to be fruitful, which is still an open problem in sociology, but a promising program to pursue (Cetina and Cicourel 2014; Coleman 1986).

Second, since the American Community Survey did not start collecting data until 2005, our analysis only includes migration-flow networks for two time points (2006-2010, 2011-2015). This

data limitation prevents us from conducting dynamic analysis about changes in intercounty migration patterns throughout the past decades, and therefore, our findings do not speak directly to the reasons behind the long-term decline of migration. Yet, our identification of drivers and especially inhibitors behind migration flows could serve as a starting point for this inquiry. For example, since political division across geographical areas deters migration, it may be worthwhile for future research to examine how the geography of politics and preference about political homophily have changed over time, and how the evolution of political landscapes and polarization relates to the long-term decline of migration. Studies of the changing patterns of immigrant inflows and the relationship between internal and international migration flows can illuminate the change of population dynamics over time. Integration of knockout experiments via network simulation and historical data about political climate and migration/immigration flows might be one approach to advance the inquiry into the social forces behind the growing immobility in the United States. In addition, future research might also benefit from exploring the changing balance of forces of the competing internal dynamics of the migration system over the past decades. Given that the VTERGM framework we employ here is capable of handling networks with multiple time steps, our analytical framework could be employed for dynamic analysis once migration-flow data for more time points becomes available.

In like vein, the time period we analyzed covers the Great Recession (Grusky et al. 2011). Despite our controls of various economic factors, it is possible that some aspects of our findings may be particular to this period, as economic shocks can influence migration patterns (Monras 2018; cf. Molloy et al. 2011). Specifically, since economic recession can suppress migration, it is possible that fewer waypoint flows are consequence of the period effect that temporarily suppresses stepwise migration. Nevertheless, the formal expressions of relational linkages and network patterns, and the modeling of migration-flow networks using ERGMs are generally applicable to study migration flows of different periods and regions at different scales. Future research may consider replicate and compare analysis of relational and network patterns of migration flows in different time and space using similar frameworks; they will reveal what patterns are context-specific in

certain spatial-temporal settings, and which are generalizable to migration in other societies.

Furthermore, another fruitful direction for future work is to complicate the analysis of internal dynamics of migration system by examining higher-order dependence structure of (valued) networks. One example is network transitivity, a structural feature associated with hierarchy within the migration system (Leal 2021). We do not observe a strong transitive hierarchical system in the U.S. internal migration system, as indicated by the lack of waypoint flows, *ceteris paribus*.[15] Nevertheless, transitivity is in general a theoretically-interesting dependence structure for study of mobility networks, and should ideally be examined in valued networks so to consider the quantitative feature of migration flows. This requires theoretical and methodological developments in formal specification of dependence terms in the valued network setting, e.g., clarifying the properties of different definitions of transitivity and their relationship to network degeneracy (Krivitsky 2012). It also demands further advancements in computational methods for valued network models to allow for evaluation of more complicated dependence structures in large networks.

Last but not least, as population immobility has become a long-term phenomenon in the U.S., it poses important questions about its broader social implications. Future research could explore the relationship between geographical mobility and social mobility, and how the divergent geographical mobility patterns across various social groups may influence their life chances and well being. A lack of population exchange, especially between localities with different cultural and political climates, could have ramifications on the social divisions of the country. Two decades ago, Putnam's (2000) *Bowling Alone* embarked the great debates about the "collapse of American communities," marked by the detachment and disengagement of individuals from their communities. Observing the population segmentation and immobility, it raises the question whether we are witnessing the "tribalization of American communities," where local communities diverge in their demographics, culture, and policy, with limited interaction, communication, and cooperation among people and organizations from dissimilar local communities.

In conclusion, grappling with the mobility bias in migration studies, this paper utilizes migration systems theory and network methods to study the mechanisms behind population immobility

in the United States. We identify segmentation as a significant feature of the American migration landscape, which has potentially immobilized millions of intercounty migration each year in the 2010s. The paper demonstrates how network and simulation methods can contribute to a systemic understanding of mobility and population dynamics. We also call for more theoretical and empirical research about the interrelationships between migration, segregation, and polarization, and how they shape the foundation of social lives in America and beyond.

Notes

[1]As an example, Eeckhout (2004:1431) contends that "the central thesis in this paper: population mobility is driven by economic forces."

[2]By valued network (or weighted network), we refer to networks whose ties are not binary (present or absent), but are associated with a quantitative value; specifically, tie values in this study indicate the volume of migration flows between directed pairs of U.S. counties.

[3]Bakewell (2010, 2014) and DeWaard and Ha (2019) have debated about whether and how studies of migration networks contribute to MST. As this paper shows, echoing Leal (2021), we agree with DeWaard and Ha (2019) that network analysis is an effective way of theorizing and testing the structures and dynamics of migration across geography; we also recognize Bakewell's critique that network analysis of migration flows is one of the many approaches to study migration systems, and that students of MST should beware the pitfall of abstract and static descriptions of migration systems. In this regard, this paper leverages theories and empirical findings in migration studies to motivate tests about structures and patterns of migration networks. We also call for more research with different levels of analysis to triangulate our findings for a comprehensive understanding of migration and immobility.

[4]We thank an anonymous reviewer for pointing out this distinction.

[5]This is because transitive hierarchy is a network structure built on waypoint flow, and an underrepresentation of the former necessarily implies an underrepresentation of the latter. It is possible that in this circumstance there can be net tendency for waypoint flows to be transitively rather than cyclically closed *where they occur*. But one will still see fewer transitive closures (as there are fewer paths to close in the first place) than one would expect by chance. Put another way, standard transitivity effects measure the overrepresentation of both waypoint flow and transitive closure, not merely the latter.

[6]Since the phrasing of "balkanization" can be construed to carry certain normative connotations regarding immigration, we follow the practice in Kritz and Gurak (2001), and phrase the phenomenon as the internal migratory

response to immigration.

[7]We also simulated networks using the full model (without knockouts), and calculated the difference in the total migrant size between the full-model simulation and the observed, as a measurement of bias introduced in the procedure. We then corrected the total population sizes in knockout scenarios by extracting that difference. As the difference is 0.7% of the observed migration volume, corrected and uncorrected estimates are nearly identical.

[8]Another dataset that reports counts of county-to-county migration flows is offered by the Internal Revenue Service (IRS) (Hauer and Byars 2019). While ACS is a nationally representative demographic survey, the representativeness is a potential concern of the IRS data, as it only contains people filing tax returns, and therefore are not representative of the elder, the low-income, and the immigrant populations. Further, the IRS data of the post 2011-2012 period currently suffers from systemic problems that are not yet resolved (DeWaard et al. 2022). Nonetheless, the IRS reports migration data annually, and can be useful for fine-grained dynamic analysis of migration before 2011.

[9]We use the L1 Euclidean distance measure, or what was called the dissimilarity score in social segregation literature (Massey and Denton 1988).

[10]Given how Hawaii and Alaska calculate their election results, we conduct the following operations to map their local election data to counties. Since Kalawao County, HI is regarded as a part of Maui County, HI for election purposes, we input the election results of both counties with their pooled results. Election results in Alaska were reported by election districts rather than counties. We used the map to match election results of the 40 districts with the 28 counties. The result of a county was input with that of its district if the county was affiliated with one single district. We take the mean of the results of the districts that a county spans if the county is affiliated with multiple districts. The approximation would underestimate the political difference between counties, but the bias should be minor as the affected county takes less than 1% of the sample. We thank the election offices of Hawaii and Alaska for clarification and maps of the election districts during 2002-2013 in Alaska.

[11]The reciprocity statistic calculates the summation of minimum value of each pair of edges by dyad. Formally, $g_m(y) = \sum_{(i,j)\in\mathbb{Y}} min(y_{ij}, y_{ji})$, where $\mathbb{Y}$ denotes the set of all i, j pairs.

[12]Formally: $g_f = \sum_{i\in\mathbb{V}} min\{\sum_{j\in\mathbb{V}, j\neq i} y_{ij}, \sum_{k\in\mathbb{V}, k\neq i} y_{ki}\}$, where $\mathbb{V}$ is the set of all vertices/nodes (counties), and y_{ij}, y_{ki} are values of the edge from county i to j and k to i, respectively. The term is similar to the 2-paths or mixed-2-stars in binary ERGMs, which is the number of times a node receives an edge and sends another (Morris et al. 2008).

[13]We note that this conclusion depends on the assumption that the context dissimilarity influences people's decision of whether to migrate or not, and not merely influencing their choice of destination. We would thus not expect this model to accurately predict involuntary migration in response to events like political turmoil or natural disasters, which dominate people's decision of migrating or not under those circumstances. However, these seem unlikely to have been significant drivers of internal migration in the U.S. during the study period. We thank the anonymous reviewer for pointing out this assumption.

[14]Since this is an aggregate-level analysis of population flows, the finding does not distinguish the characteristics of internal migrants, such as their race and ethnicity or socioeconomic status. Hence, we do not directly engage with more fine-grained debates about whether immigration deters in-migration and promotes out-migration of certain population categories as predicted by some literature (Frey 1995a), which requires more detailed data.

[15]As discussed in Hypotheses, both waypoint flow and transitivity are triadic features that concern edge structure in an (i, j, k) triple; waypoint flow captures the"backbone" of flow within the triple ($i \to j \to k$), while transitive triads involve the co-presence of waypoint flow and a direct $i \to k$ flow. The negative effect for waypoint flow in our models means that triples with strong $i \to j \to k$ paths are suppressed, which also necessarily suppresses transitive triples net of other effects in the model. Interestingly, while the waypoint flow (and its binary-network version, two-paths) is a more basic lower-level dependence structure, which carries motivations from social behavior patterns such as those detailed in this paper, it receives relatively less examination in the network literature. We hope this paper helps draw more attention to waypoint flow and other triadic network structures of potential substantive importance for flow networks.

References

Almquist, Zack W. and Carter T. Butts. 2015. "Predicting Regional Self-Identification from Spatial Network Models." *Geographical Analysis* 47:50–72.

Badger, Emily, Quoctrung Bui, and Josh Katz. 2018. "The Suburbs Are Changing. But Not in All the Ways Liberals Hope." *The New York Times* .

Bakewell, Oliver. 2010. "Some Reflections on Structure and Agency in Migration Theory." *Journal of Ethnic and Migration Studies* 36:1689–1708.

Bakewell, Oliver. 2014. "Relaunching migration systems." *Migration Studies* 2:300–318.

Bakewell, Oliver, Hein De Haas, and Agnieszka Kubal. 2012. "Migration Systems, Pioneer Migrants and the Role of Agency." *Journal of Critical Realism* 11:413–437.

Bakewell, Oliver, Godfried Engbersen, Maria Lucinda Fonseca, and Cindy Horst. 2016a. *Beyond Networks: Feedback in International Migration*. Springer.

Bakewell, Oliver, Agnieszka Kubal, and Sónia Pereira. 2016b. "Introduction: Feedback in Migration Processes." In *Beyond Networks: Feedback in International Migration*, edited by Oliver Bakewell, Godfried Engbersen, Maria Lucinda Fonseca, and Cindy Horst, Migration, Diasporas and Citizenship, pp. 1–17. London: Palgrave Macmillan UK.

Baldassarri, Delia and Andrew Gelman. 2008. "Partisans without Constraint: Political Polarization and Trends in American Public Opinion." *American Journal of Sociology* 114:408–446.

Bauman, Zygmunt. 2000. *Liquid Modernity*. John Wiley & Sons.

Bishop, Bill and Robert G. Cushing. 2009. *The Big Sort: Why the Clustering of Like-minded America is Tearing Us Apart*. Houghton Mifflin Harcourt.

Borjas, George J. 2006. "Native Internal Migration and the Labor Market Impact of Immigration." *Journal of Human Resources* 41:221–258.

Boyle, Paul, Halfacree Keith H., Robinson Vaughan, and Robinson Vaughan. 2014. *Exploring Contemporary Migration*. Abingdon, United Kingdom: Routledge.

Briggs, Xavier de Souza, Susan J. Popkin, and John Goering. 2010. *Moving to Opportunity: The Story of an American Experiment to Fight Ghetto Poverty*. Oxford University Press.

Brown, Jacob R. and Ryan D. Enos. 2021. "The measurement of partisan sorting for 180 million voters." *Nature Human Behaviour* 5:998–1008.

Brown, Susan K. and Frank D. Bean. 2016. "Conceptualizing Migration: From Internal/International to Kinds of Membership." In *International Handbook of Migration and Population Distribution*, edited by Michael J. White, International Handbooks of Population, pp. 91–106. Dordrecht: Springer Netherlands.

Burris, Val. 2004. "The Academic Caste System: Prestige Hierarchies in PhD Exchange Networks." *American Sociological Review* 69:239–264.

Butts, Carter T. 2007. "Models for Generalized Location Systems." *Sociological Methodology* 37:283–348.

Butts, Carter T., Ryan M. Acton, John R. Hipp, and Nicholas N. Nagle. 2012. "Geographical variability and network structure." *Social Networks* 34:82–100.

Card, David. 2001. "Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration." *Journal of Labor Economics* 19:22–64.

Carling, Jørgen. 2002. "Migration in the age of involuntary immobility: Theoretical reflections and Cape Verdean experiences." *Journal of Ethnic and Migration Studies* 28:5–42.

Carling, Jørgen and Kerilyn Schewel. 2018. "Revisiting aspiration and ability in international migration." *Journal of Ethnic and Migration Studies* 44:945–963.

Castles, Stephen, Hein de Haas, and Mark J. Miller. 2013. *The Age of Migration: International Population Movements in the Modern World*. Palgrave Macmillan.

Cetina, Karin Knorr and A. V. Cicourel (eds.). 2014. *Advances in Social Theory and Methodology (RLE Social Theory): Toward an Integration of Micro- and Macro-Sociologies*. London: Routledge.

Charyyev, Batyr and Mehmet Hadi Gunes. 2019. "Complex network of United States migration." *Computational Social Networks* 6:1–28.

Chen, M. Keith and Ryne Rohla. 2018. "The effect of partisanship and political advertising on close family ties." *Science* 360:1020–1024.

Cheng, Siwei and Barum Park. 2020. "Flows and Boundaries: A Network Approach to Studying Occupational Mobility in the Labor Market." *American Journal of Sociology* 126:577–631.

Clark, William A. V. 2008. "Reexamining the moving to opportunity study and its contribution to changing the distribution of poverty and ethnic concentration." *Demography* 45:515–535.

Cohen, Joel E., Marta Roig, Daniel C. Reuman, and Cai GoGwilt. 2008. "International migration beyond gravity: A statistical model for use in population projections." *Proceedings of the National Academy of Sciences* 105:15269–15274.

Cohen, Jeffrey H. and Ibrahim Sirkeci. 2011. *Cultures of Migration: The Global Nature of Contemporary Mobility*. University of Texas Press.

Coleman, James S. 1986. "Social Theory, Social Research, and a Theory of Action." *American Journal of Sociology* 91:1309–1335.

Conway, Dennis. 1980. "Step-Wise Migration: Toward a Clarification of the Mechanism." *International Migration Review* 14:3–14.

Cooke, Thomas J. 2013. "Internal Migration in Decline." *The Professional Geographer* 65:664–675.

Coulson, N. Edward and Paul L. E. Grieco. 2013. "Mobility and mortgages: Evidence from the PSID." *Regional Science and Urban Economics* 43:1–7.

Cramer, Katherine J. 2016. *The Politics of Resentment: Rural Consciousness in Wisconsin and the Rise of Scott Walker*. Chicago: University of Chicago Press.

Crowder, Kyle, Jeremy Pais, and Scott J. South. 2012. "Neighborhood Diversity, Metropolitan Constraints, and Household Migration." *American Sociological Review* 77:325–353.

Czaika, Mathias and Hein de Haas. 2017. "The Effect of Visas on Migration Processes." *International Migration Review* 51:893–926.

Dao, Mai, Davide Furceri, and Prakash Loungani. 2017. "Regional Labor Market Adjustment in the United States: Trend and Cycle." *The Review of Economics and Statistics* 99:243–257.

Davis, James A and Samuel Leinhardt. 1972. "The Structure of Positive Interpersonal Relations in Small Groups." In *Sociological Theories in Progress*, pp. 218–251. Boston: Houghton Mifflin.

de Haas, Hein. 2010. "The Internal Dynamics of Migration Processes: A Theoretical Inquiry." *Journal of Ethnic and Migration Studies* 36:1587–1617.

DellaPosta, Daniel. 2020. "Pluralistic Collapse: The "Oil Spill" Model of Mass Opinion Polarization." *American Sociological Review* 85:507–536.

DellaPosta, Daniel and Michael Macy. 2015. "The center cannot hold. Networks, echo chambers, and polarization." In *Order on the Edge of Chaos, Social Psychology and the Problem of Social Order*, pp. 86–104. New York: Cambridge University Press.

DellaPosta, Daniel, Yongren Shi, and Michael Macy. 2015. "Why Do Liberals Drink Lattes?" *American Journal of Sociology* 120:1473–1511.

DeLuca, Stefanie, Holly Wood, and Peter Rosenblatt. 2019. "Why Poor Families Move (And Where They Go): Reactive Mobility and Residential Decisions." *City & Community* 18:556–593.

Desmarais, Bruce A. and Skyler J. Cranmer. 2012. "Statistical Inference for Valued-Edge Networks: The Generalized Exponential Random Graph Model." *PLoS ONE* 7:e30136.

DeWaard, Jack, Katherine J. Curtis, and Elizabeth Fussell. 2016. "Population recovery in New Orleans after Hurricane Katrina: exploring the potential role of stage migration in migration systems." *Population and Environment* 37:449–463.

DeWaard, Jack, Elizabeth Fussell, Katherine J. Curtis, and Jasmine Trang Ha. 2020. "Changing spatial interconnectivity during the "Great American Migration Slowdown": A decomposition of intercounty migration rates, 1990–2010." *Population, Space and Place* 26:e2274.

DeWaard, Jack and Jasmine Trang Ha. 2019. "Resituating relaunched migration systems as emergent entities manifested in geographic structures." *Migration Studies* 7:39–58.

DeWaard, Jack, Mathew Hauer, Elizabeth Fussell, Katherine J. Curtis, Stephan D. Whitaker, Kathryn McConnell, Kobie Price, David Egan-Robertson, Michael Soto, and Catalina Anampa Castro. 2022. "User Beware: Concerning Findings from the Post 2011–2012 U.S. Internal Revenue Service Migration Data." *Population Research and Policy Review* 41:437–448.

DeWaard, Jack, Keuntae Kim, and James Raymer. 2012. "Migration Systems in Europe: Evidence From Harmonized Flow Data." *Demography* 49:1307–1333.

DiPrete, Thomas A., Andrew Gelman, Tyler McCormick, Julien Teitler, and Tian Zheng. 2011. "Segregation in Social Networks Based on Acquaintanceship and Trust." *American Journal of Sociology* 116:1234–83.

Durand, Jorge and Douglas S. Massey. 2010. "New world orders: Continuities and changes in Latin American migration." *The Annals of the American Academy of Political and Social Science* 630:20–52. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Dye, Thomas R. 1990. *American Federalism: Competition Among Governments*. Lexington Books.

Eeckhout, Jan. 2004. "Gibrat's Law for (All) Cities." *American Economic Review* 94:1429–1451.

Einstein, A., B. Podolsky, and N. Rosen. 1935. "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?" *Physical Review* 47:777–780. Cambridge, UK: Cambridge University Press.

Fawcett, James T. 1989. "Networks, Linkages, and Migration Systems." *The International Migration Review* 23:671–680.

Fernández-Kelly, Patricia. 2008. "The Back Pocket Map: Social Class and Cultural Capital as Transferable Assets in the Advancement of Second-Generation Immigrants." *The ANNALS of the American Academy of Political and Social Science* 620:116–137.

Fischer, Claude S. 2002. "Ever-More Rooted Americans." *City & Community* 1:177–198.

Fitchen, Janet M. 1994. "Residential Mobility Among the Rural Poor." *Rural Sociology* 59:416–436. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1549-0831.1994.tb00540.x.

Fossett, Mark. 2006. "Ethnic Preferences, Social Distance Dynamics, and Residential Segregation: Theoretical Explorations Using Simulation Analysis." *The Journal of Mathematical Sociology* 30:185–273.

Freier, Luisa Feline and Kyle Holloway. 2019. "The Impact of Tourist Visas on Intercontinental South-South Migration: Ecuador's Policy of "Open Doors" as a Quasi-Experiment." *International Migration Review* 53:1171–1208.

Frey, William H. 1995a. "Immigration and Internal Migration 'Flight' from US Metropolitan Areas: Toward a New Demographic Balkanisation." *Urban Studies* 32:733–757.

Frey, William H. 1995b. "Immigration and internal migration "flight": A California case study." *Population and Environment* 16:353–375.

Frey, William H. 2009. "The great American migration slowdown: Regional and metropolitan dimensions." Technical report, Brookings Institution, Washington, DC.

Garip, Filiz. 2008. "Social capital and migration: How do similar resources lead to divergent outcomes?" *Demography* 45:591–617.

Garip, Filiz and Asad L. Asad. 2016. "Network Effects in Mexico–U.S. Migration: Disentangling the Underlying Social Mechanisms." *American Behavioral Scientist* 60:1168–1193.

Gimpel, James G. and Iris S. Hui. 2015. "Seeking politically compatible neighbors? The role of neighborhood partisan composition in residential sorting." *Political Geography* 48:130–142.

Gondal, Neha. 2018. "Duality of departmental specializations and PhD exchange: A Weberian analysis of status in interaction using multilevel exponential random graph models (mERGM)." *Social Networks* 55:202–212.

Goodreau, Steven M., James A. Kitts, and Martina Morris. 2009. "Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks." *Demography* 46:103–125.

Grusky, David B., Bruce Western, and Christopher Wimer. 2011. *The Great Recession*. Russell Sage Foundation.

Hall, Bradford, Advait Limaye, and Ashok B. Kulkarni. 2009. "Overview: Generation of Gene Knockout Mice." *Current Protocols in Cell Biology* 44:19.12.1–19.12.17.

Han, Xiaoyi, Yilan Xu, Linlin Fan, Yi Huang, Minhong Xu, and Song Gao. 2021. "Quantifying COVID-19 importation risk in a dynamic network of domestic cities and international countries." *Proceedings of the National Academy of Sciences* 118:e2100201118.

Hauer, Mathew and James Byars. 2019. "IRS county-to-county migration data, 1990–2010." *Demographic Research* 40:1153–1166.

Hauer, Mathew E. 2017. "Migration induced by sea-level rise could reshape the US population landscape." *Nature Climate Change* 7:321–325.

Herting, Jerald R., David B. Grusky, and Stephen E. Van Rompaey. 1997. "The Social Geography of Interstate Mobility and Persistence." *American Sociological Review* 62:267–287.

Hipp, John R., Carter T. Butts, Ryan Acton, Nicholas N. Nagle, and Adam Boessen. 2013. "Extrapolative simulation of neighborhood networks based on population spatial distribution: Do they predict crime?" *Social Networks* 35:614–625.

Hochschild, Arlie Russell. 2018. *Strangers in Their Own Land: Anger and Mourning on the American Right*. The New Press.

Huang, Peng and Carter T. Butts. 2024. "Parameter estimation procedures for exponential-family random graph models on count-valued networks: A comparative simulation study." *Social Networks* 76:51–67.

Hunter, David R., Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. 2008. "ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of statistical software* 24:nihpa54860.

Hyatt, Henry, Erika McEntarfer, Ken Ueda, and Alexandria Zhang. 2018. "Interstate Migration and Employer-to-Employer Transitions in the United States: New Evidence From Administrative Records Data." *Demography* 55:2161–2180.

Hyatt, Henry R. and James R. Spletzer. 2013. "The recent decline in employment dynamics." *IZA Journal of Labor Economics* 2:1–21.

Intrator, Jake, Jonathan Tannen, and Douglas S. Massey. 2016. "Segregation by race and income in the United States 1970–2010." *Social Science Research* 60:45–60.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. 2019. "The Origins and Consequences of Affective Polarization in the United States." *Annual Review of Political Science* 22:129–146.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, Not IdeologyA Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76:405–431.

Jasso, Guillermina. 2011. "Migration and stratification." *Social Science Research* 40:1292–1336.

Jennissen, Roel. 2007. "Causality Chains in the International Migration Systems Approach." *Population Research and Policy Review* 26:411–436.

Jia, Ning, Raven Molloy, Christopher L. Smith, and Abigail Wozniak. 2022. "The Economics of Internal Migration: Advances and Policy Questions." Finance and Economics Discussion Series 2022-003. Washington: Board of Governors of the Federal Reserve System.

Johnson, Janna E. and Morris M. Kleiner. 2020. "Is Occupational Licensing a Barrier to Interstate Migration?" *American Economic Journal: Economic Policy* 12:347–373.

Kaplan, Greg and Sam Schulhofer-Wohl. 2017. "Understanding the Long-Run Decline in Interstate Migration." *International Economic Review* 58:57–94.

Kim, Keuntae and Joel E. Cohen. 2010. "Determinants of International Migration Flows to and from Industrialized Countries: A Panel Data Approach beyond Gravity." *International Migration Review* 44:899–932.

Krackhardt, David and Lyman W. Porter. 1986. "The snowball effect: Turnover embedded in communication networks." *Journal of Applied Psychology* 71:50–55.

Kreager, Derek A., Jacob T.N. Young, Dana L. Haynie, Martin Bouchard, David R. Schaefer, and Gary Zajac. 2017. "Where "Old Heads" Prevail: Inmate Hierarchy in a Men's Prison Unit." *American Sociological Review* 82:685–718.

Kritz, Mary M and Douglas T Gurak. 2001. "The impact of immigration on the internal migration of natives and immigrants." *Demography* 38:133–145.

Kritz, Mary M., Lin Lean Lim, Hania Zlotnik, and Lin Lean Lin Lean Lim. 1992. *International migration systems: a global approach*. Oxford University Press, USA.

Krivitsky, Pavel N. 2012. "Exponential-family random graph models for valued networks." *Electronic journal of statistics* 6:1100–1128.

Krivitsky, Pavel N and Carter T Butts. 2013. "Modeling valued networks with statnet." *The Statnet Development Team* .

Krysan, Maria and Kyle Crowder. 2017. *Cycle of Segregation: Social Processes and Residential Stratification*. Russell Sage Foundation.

Lakon, Cynthia M., John R. Hipp, Cheng Wang, Carter T. Butts, and Rupa Jose. 2015. "Simulating Dynamic Network Models and Adolescent Smoking: The Impact of Varying Peer Influence and Peer Selection." *American Journal of Public Health* 105:2438–2448.

Leal, Diego F. 2021. "Network Inequalities and International Migration in the Americas." *American Journal of Sociology* 126:1067–1126.

Lee, Everett S. 1966. "A theory of migration." *Demography* 3:47–57.

Lee, Jennifer and Min Zhou. 2017. "Why class matters less for Asian-American academic achievement." *Journal of Ethnic and Migration Studies* 43:2316–2330.

Leszczensky, Lars and Sebastian Pink. 2019. "What Drives Ethnic Homophily? A Relational Approach on How Ethnic Identification Moderates Preferences for Same-Ethnic Friends." *American Sociological Review* 84:394–419.

Levendusky, Matthew. 2009. *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans*. University of Chicago Press.

Lewis, Kevin. 2013. "The limits of racial prejudice." *Proceedings of the National Academy of Sciences* 110:18814–18819.

Lewis, Kevin. 2016. "Preferences in the Early Stages of Mate Choice." *Social Forces* 95:283–320.

Lewis, Kevin and Andrew V. Papachristos. 2019. "Rules of the Game: Exponential Random Graph Models of a Gang Homicide Network." *Social Forces* 98:1829–1858.

Liang, Zai, Miao David Chunyu, Guotu Zhuang, and Wenzhen Ye. 2008. "Cumulative Causation, Market Transition, and Emigration from China." *American Journal of Sociology* 114:706–737.

Liang, Zai and Miao David Chunyu. 2013. "Migration within China and from China to the USA: The effects of migration networks, selectivity, and the rural political economy in Fujian Province." *Population Studies* 67:209–223.

Lichter, Daniel T., Domenico Parisi, and Michael C. Taquino. 2022. "Inter-County Migration and the Spatial Concentration of Poverty: Comparing Metro and Nonmetro Patterns." *Rural Sociology* 87:119–143.

Liu, Xi, Clio Andris, and Bruce A. Desmarais. 2019. "Migration and political polarization in the U.S.: An analysis of the county-level migration network." *PLOS ONE* 14:e0225405.

Lobao, Linda and Paige Kelly. 2019. "Local Governments across the Rural–Urban Continuum: Findings from a Recent National County Government Study." *State and Local Government Review* 51:223–232.

Long, Larry. 1991. "Residential Mobility Differences among Developed Countries." *International Regional Science Review* 14:133–147.

Lu, Yao, Zai Liang, and Miao David Chunyu. 2013. "Emigration from China in Comparative Perspective." *Social Forces* 92:631–658.

Lubbers, Miranda Jessica, Ashton M. Verdery, and José Luis Molina. 2020. "Social Networks and Transnational Social Fields: A Review of Quantitative and Mixed-Methods Approaches." *International Migration Review* 54:177–204.

Mabogunje, Akin L. 1970. "Systems Approach to a Theory of Rural-Urban Migration." *Geographical Analysis* 2:1–18.

Massey, D.S. and N.A. Denton. 1988. "The dimensions of residential segregation." *Social Forces* 67:281–315.

Massey, Douglas and Nancy A. Denton. 1993. *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press.

Massey, Douglas S. 1990. "Social Structure, Household Strategies, and the Cumulative Causation of Migration." *Population Index* 56:3.

Massey, Douglas S., Joaquin Arango, Graeme Hugo, Ali Kouaouci, and Adela Pellegrino. 1999. *Worlds in Motion: Understanding International Migration at the End of the Millennium*. Clarendon Press.

Massey, Douglas S., Joaquin Arango, Graeme Hugo, Ali Kouaouci, Adela Pellegrino, and J. Edward Taylor. 1993. "Theories of International Migration: A Review and Appraisal." *Population and Development Review* 19:431–466.

Massey, Douglas S., Jorge Durand, and Karen A. Pren. 2016. "Why Border Enforcement Backfired." *American Journal of Sociology* 121:1557–1600.

Massey, Douglas S. and Kristin E. Espinosa. 1997. "What's Driving Mexico-U.S. Migration? A Theoretical, Empirical, and Policy Analysis." *American Journal of Sociology* 102:939–999.

Massey, Douglas S., Luin Goldring, and Jorge Durand. 1994a. "Continuities in Transnational Migration: An Analysis of Nineteen Mexican Communities." *American Journal of Sociology* 99:1492–1533.

Massey, Douglas S., Andrew B. Gross, and Kumiko Shibuya. 1994b. "Migration, Segregation, and the Geographic Concentration of Poverty." *American Sociological Review* 59:425.

Massey, Douglas S. and Jonathan Tannen. 2018. "Suburbanization and segregation in the United States: 1970–2010." *Ethnic and Racial Studies* 41:1594–1611.

Mata-Codesal, Diana. 2015. "Ways of Staying Put in Ecuador: Social and Embodied Experiences of Mobility–Immobility Interactions." *Journal of Ethnic and Migration Studies* 41:2274–2290.

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

McMahan, Peter and Daniel A. McFarland. 2021. "Creative Destruction: The Structural Consequences of Scientific Curation." *American Sociological Review* 86:341–376.

McMillan, Cassie. 2019. "Tied Together: Adolescent Friendship Networks, Immigrant Status, and Health Outcomes." *Demography* 56:1075–1103.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415–444.

Miller, John H. and Scott Page. 2009. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton University Press. Publication Title: Complex Adaptive Systems.

MIT Election Data and Science Lab. 2018. "County Presidential Election Returns 2000-2016." Retrieved September 29, 2020 (https://doi.org/10.7910/DVN/VOQCHQ).

Molloy, Raven, Christopher L Smith, and Abigail Wozniak. 2011. "Internal Migration in the United States." *Journal of Economic Perspectives* 25:173–196.

Molloy, Raven, Christopher L. Smith, and Abigail Wozniak. 2017. "Job Changing and the Decline in Long-Distance Migration in the United States." *Demography* 54:631–653.

Monras, Joan. 2018. "Economic Shocks and Internal Migration." *CEPR Discussion Paper No. DP12977* .

Moody, James. 2001. "Race, School Integration, and Friendship Segregation in America." *American Journal of Sociology* 107:679–716.

Morris, Martina, Mark S. Handcock, and David R. Hunter. 2008. "Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects." *Journal of statistical software* 24:1548–7660.

Mouw, Ted, Sergio Chavez, Heather Edelblute, and Ashton Verdery. 2014. "Binational Social Networks and Assimilation: A Test of the Importance of Transnationalism." *Social Problems* 61:329–359.

Mueller, J. Tom and Stephen Gasteyer. 2023. "The ethnically and racially uneven role of water infrastructure spending in rural economic development." *Nature Water* 1:74–82. Number: 1 Publisher: Nature Publishing Group.

Mummolo, Jonathan and Clayton Nall. 2016. "Why Partisans Do Not Sort: The Constraints on Political Segregation." *The Journal of Politics* 79:45–59. Publisher: The University of Chicago Press.

National Bureau of Economic Research. 2016. "County Distance Database." Retrieved September 29, 2020 (https://data.nber.org/data/county-distance-database.html).

Nogle, June Marie. 1994. "The Systems Approach to International Migration: An Application of Network Analysis Methods." *International Migration* 32:329–342.

Palloni, Alberto, Douglas S. Massey, Miguel Ceballos, Kristin Espinosa, and Michael Spittel. 2001. "Social Capital and International Migration: A Test Using Information on Family Networks." *American Journal of Sociology* 106:1262–1298.

Papachristos, Andrew V., David M. Hureau, and Anthony A. Braga. 2013. "The Corner and the Crew: The Influence of Geography and Social Networks on Gang Violence." *American Sociological Review* 78:417–447.

Partridge, Mark D., Dan S. Rickman, M. Rose Olfert, and Kamar Ali. 2012. "Dwindling U.S. internal migration: Evidence of spatial equilibrium or structural shifts in local labor markets?" *Regional Science and Urban Economics* 42:375–388.

Paul, Anju Mary. 2011. "Stepwise International Migration: A Multistage Migration Pattern for the Aspiring Migrant." *American Journal of Sociology* 116:1842–86.

Paul, Anju Mary. 2017. *Multinational Maids: Stepwise Migration in a Global Labor Market*. Cambridge University Press.

Plantinga, Andrew J., Cécile Détang-Dessendre, Gary L. Hunt, and Virginie Piguet. 2013. "Housing prices and inter-urban migration." *Regional Science and Urban Economics* 43:296–306.

Poot, Jacques, Omoniyi Alimi, Michael P. Cameron, and David C. Maré. 2016. "The Gravity Model of Migration: The Successful Comeback of an Ageing Superstar in Regional Science." *SSRN Electronic Journal* .

Preuhs, Robert R. 1999. "State Policy Components of Interstate Migration in the United States." *Political Research Quarterly* 52:527–549.

Preuhs, Robert R. 2020. "Pack Your Politics! Assessing the Vote Choice of Latino Interstate Migrants." *The Journal of Race, Ethnicity, and Politics* 5:130–165.

Putnam, Robert D. 2000. *Bowling Alone: The Collapse and Revival of American Community*. Simon and Schuster.

Quillian, Lincoln. 2015. "A Comparison of Traditional and Discrete-Choice Approaches to the Analysis of Residential Mobility and Locational Attainment." *The ANNALS of the American Academy of Political and Social Science* 660:240–260.

Ravenstein, E. G. 1885. "The Laws of Migration." *Journal of the Royal Statistical Society* 48:167–235.

Raymer, James and Andrei Rogers. 2007. "Applying Model Migration Schedules to Represent Age-Specific Migration Flows." In *International Migration in Europe*, edited by James Raymer and Frans Willekens, pp. 175–192. Chichester, UK: John Wiley & Sons, Ltd.

Riddell, J. Barry and Milton E. Harvey. 1972. "The Urban System in the Migration Process: An Evaluation of Step-Wise Migration in Sierra Leone." *Economic Geography* 48:270.

Rogers, Andrei and Luis J Castro. 1981. *Model Migration Schedules*. Laxenburg: International Institute for Applied Systems Analysis.

Ryo, Emily. 2013. "Deciding to Cross: Norms and Economics of Unauthorized Migration." *American Sociological Review* 78:574–603.

Sakoda, James M. 1971. "The checkerboard model of social interaction." *The Journal of Mathematical Sociology* 1:119–132.

Schelling, Thomas C. 1969. "Models of Segregation." *American Economic Review* 59:483–493.

Schelling, Thomas C. 1978. *Micromotives and Macrobehavior*. W. W. Norton & Company.

Schewel, Kerilyn. 2020. "Understanding Immobility: Moving Beyond the Mobility Bias in Migration Studies." *International Migration Review* 54:328–355.

Schildkraut, Deborah J, Tomás R Jiménez, John F Dovidio, and Yuen J Huo. 2019. "A Tale of Two States: How State Immigration Climate Affects Belonging to State and Country among Latinos." *Social Problems* 66:332–355.

Schiller, Nina Glick, Linda Basch, and Cristina Szanton Blanc. 1995. "From Immigrant to Transmigrant: Theorizing Transnational Migration." *Anthropological Quarterly* 68:48–63.

Schroeder, Jonathan P. and José D. Pacas. 2021. "Across the Rural–Urban Universe: Two Continuous Indices of Urbanization for U.S. Census Microdata." *Spatial Demography* 9:131–154.

Sharkey, Patrick. 2015. "Geographic Migration of Black and White Families Over Four Generations." *Demography* 52:209–231.

Smith, Chris M. and Andrew V. Papachristos. 2016. "Trust Thy Crooked Neighbor: Multiplexity in Chicago Organized Crime Networks." *American Sociological Review* 81:644–667.

Smith, Jeffrey A., Miller McPherson, and Lynn Smith-Lovin. 2014. "Social Distance in the United States: Sex, Race, Religion, Age, and Education Homophily among Confidants, 1985 to 2004." *American Sociological Review* 79:432–456.

Sparrowe, Raymond T. and Robert C. Liden. 1997. "Process and Structure in Leader-Member Exchange." *Academy of Management Review* 22:522–552.

Spring, Amy, Clara H. Mulder, Michael J. Thomas, and Thomas J. Cooke. 2021. "Migration after union dissolution in the United States: The role of non-resident family." *Social Science Research* 96:102539.

Srivastava, Sameer B. and Mahzarin R. Banaji. 2011. "Culture, Cognition, and Collaborative Networks in Organizations." *American Sociological Review* 76:207–233.

Steinbeck, John. 1939. *The Grapes of Wrath*. Penguin.

Stockdale, Aileen and Tialda Haartsen. 2018. "Editorial introduction: Putting rural stayers in the spotlight." *Population, Space and Place* 24:e2124.

Tam Cho, Wendy K., James G. Gimpel, and Iris S. Hui. 2013. "Voter Migration and the Geographic Sorting of the American Electorate." *Annals of the Association of American Geographers* 103:856–870.

Thomas, Loring J., Peng Huang, Fan Yin, Junlan Xu, Zack W. Almquist, John R. Hipp, and Carter T. Butts. 2022. "Geographical patterns of social cohesion drive disparities in early COVID infection hazard." *Proceedings of the National Academy of Sciences* 119:e2121675119.

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

Tocqueville, Alexis. 1834. *Democracy in America: And Two Essays on America*. Penguin UK.

Todaro, Michael P. 1976. *Internal migration in developing countries*. Genève, Switzerland: International Labour Office.

Tolnay, Stewart E. 2003. "The African American "Great Migration" and Beyond." *Annual Review of Sociology* 29:209–232.

Treyz, George I., Dan S. Rickman, Gary L. Hunt, and Michael J. Greenwood. 1993. "The Dynamics of U.S. Internal Migration." *The Review of Economics and Statistics* 75:209–214.

U.S. Census Bureau. 2013. "Census Regions and Divisions of the United States." Retrieved September 29, 2020 (https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf).

Verdery, Ashton M., Ted Mouw, Heather Edelblute, and Sergio Chavez. 2018. "Communication flows and the durability of a transnational social field." *Social Networks* 53:57–71.

Vogel, Gretchen. 2007. "A Knockout Award in Medicine." *Science* 318:178–179.

von Reichert, Christiane, John B. Cromartie, and Ryan O. Arthun. 2014a. "Impacts of Return Migration on Rural U.S. Communities." *Rural Sociology* 79:200–226.

von Reichert, Christiane, John B. Cromartie, and Ryan O. Arthun. 2014b. "Reasons for Returning and Not Returning to Rural U.S. Communities." *The Professional Geographer* 66:58–72.

Vögtle, EvaMaria and Michael Windzio. 2022. "The 'Global South' in the transnational student mobility network. Effects of institutional instability, reputation of the higher education systems, post-colonial ties, and culture." *Globalisation, Societies and Education* pp. 1–19.

Waldinger, Roger. 2013. "Immigrant transnationalism." *Current Sociology* 61:756–777.

Waldinger, Roger and David Fitzgerald. 2004. "Transnationalism in Question." *American Journal of Sociology* 109:1177–1195. Publisher: The University of Chicago Press.

Wallace, Tim and Krishna Karra. 2020. "The True Colors of America's Political Spectrum Are Gray and Green." *The New York Times* .

Wallerstein, Immanuel. 2011. *The Modern World-System I: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. University of California Press. Google-Books-ID: JZqhKZ9ucc0C.

Wasserman, Stanley and Philippa Pattison. 1996. "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*." *Psychometrika* 61:401–425.

Weber, Max. 1922. *Wirtschaft und Gesellschaft: Grundriß der Verstehenden Soziologie*. Duncker & Humblot.

White, Michael J. and Zai Liang. 1998. "The effect of immigration on the internal migration of the native-born population, 1981–1990." *Population Research and Policy Review* 17:141–166.

Wilson, David Sloan, Daniel Tumminelli O'Brien, and Artura Sesma. 2009. "Human prosociality from an evolutionary perspective: variation and correlations at a city-wide scale." *Evolution and Human Behavior* 30:190–200.

Wimmer, Andreas and Kevin Lewis. 2010. "Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook." *American Journal of Sociology* 116:583–642.

Windzio, Michael. 2018. "The network of global migration 1990–2013." *Social Networks* 53:20–29.

Windzio, Michael, Céline Teney, and Sven Lenkewitz. 2019. "A network analysis of intra-EU migration flows: how regulatory policies, economic inequalities and the network-topology shape the intra-EU migration space." *Journal of Ethnic and Migration Studies* 47:951–969.

Wright, Richard A., Mark Ellis, and Michael Reibel. 1997. "The Linkage between Immigration and Internal Migration in Large Metropolitan Areas in the United States." *Economic Geography* 73:234–254.

Xie, Yu and Chunni Zhang. 2019. "The long-term impact of the Communist Revolution on social stratification in contemporary China." *Proceedings of the National Academy of Sciences* 116:19392–19397.

Zhou, Min. 1992. *Chinatown: The Socioeconomic Potential of an Urban Enclave*. Temple University Press.

Zipf, George Kingsley. 1946. "The P1 P2/D Hypothesis: On the Intercity Movement of Persons." *American Sociological Review* 11:677–686.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Oxford, England: Addison-Wesley Press.

California Exodus? A Network Model of Population Redistribution in the United States*

Peng Huang† Carter T. Butts†‡

Abstract

Motivated by debates about California's net migration loss, we employ valued exponential-family random graph models to analyze the inter-county migration flow networks in the United States. We introduce a protocol that visualizes the complex effects of potential underlying mechanisms, and perform *in silico* knockout experiments to quantify their contribution to the California Exodus. We find that racial dynamics contribute to the California Exodus, urbanization ameliorates it, and political climate and housing costs have little impact. Moreover, the severity of the California Exodus depends on how one measures it, and California is not the state with the most substantial population loss. The paper demonstrates how generative statistical models can provide mechanistic insights beyond simple hypothesis-testing.

Keywords: Migration, Population redistribution, Valued networks, ERGMs, Simulation

1 Introduction

The "California Exodus" - a putative phenomenon in which large numbers of individuals are allegedly leaving California and migrating to other U.S. states, has become an increasingly common topic in public discourse surrounding migration and policy in the United States (e.g. Bahnsen, 2021; Beam, 2021; Dorsey, 2021; Hiltzik, 2020; Song, 2021). Popularized within conservative media circles (Bahnsen, 2021; Dorsey, 2021), the notion of a "California Exodus" serves as a focal point for a political narrative in which the state of California exemplifies the failure of the ruling Democratic party governance, and its associated social and policy regimes. Despite this politicized narrative, the net loss of California population via domestic migration *is* a long-term phenomenon, well-documented in demographic data. Nor is this a recent development: contrary to popular impression, California's net migration rate has been negative since 1989 (Hiltzik, 2020). The migration pattern of America's most populous state illuminates important trends of population redistribution in the United States, and could potentially shift the country's economic and political landscape. Historically, internal migration has played a key role in shaping the spatial distribution of population, with the most well-known and general example being urbanization (Ravenstein, 1885). In the U.S., internal migration has also played a critical role in its demographic change, including the great migration of African Americans from the South to the North (Tolnay, 2003), the westward shift of population towards the Pacific coast (Plane, 1999), and the ex-urbanization process (Plane et al., 2005).

Yet, compared to its intense treatment in popular discourse, the California Exodus as a real and persistent (if less dramatic) phenomenon receives scant attention in scientific research (c.f. Henrie and Plane, 2008).

*In press in *Journal of Mathematical Sociology*

†Departments of Sociology, and Statistics, University of California, Irvine

‡Departments of Computer Science, and EECS, University of California, Irvine

Arguably, this may be in part due to the difficulty of modeling the complexity of internal migration systems, which requires incorporating a wide range of factors influencing migration. Moreover, as migration systems theory contends (Bakewell, 2014; de Haas, 2010; Mabogunje, 1970), the migration system has endogenous feedback mechanisms, where migration flows are interdependent to each other. This further complicates mathematical models of migration flows - and their calibration to empirical data - requiring them to account for the autocorrelation structure of the system.

In this paper, we use recently developed generative network models of the internal migration system in the U.S. to help unravel the mechanisms sustaining the California Exodus, with an eye to identifying factors that may or may not contribute to this feature of the current U.S. migration system. We model the U.S. internal migration system as a network comprising counties (nodes/vertices) and migration flows between each directed pairs of counties (edges). Compared to the conventional approach that considers places as analytical units, the relational approach takes migration flows between places as units of analysis, which allows us to capture how the characteristics of origin and destination *jointly* influence their migration flows, such as the difference in political climates and costs of living. The systemic view also considers the endogenous feedback mechanism of the migration system (de Haas, 2010), reflected by the interdependence among migration flows, which gives the system its own momentum, strengthening or ameliorating the exogenous effects from the economic or political landscapes. This is achieved by specifying the network dependence structure, which accounts for the autocorrelation pattern among migration flows. The network models thus can reveal how demographic, economic, political, and geographical characteristics, together with the endogenous feedback mechanisms, shape the direction and magnitude of internal migration flows in the United States.

While computational and statistical constraints have traditionally limited network models of migration to dichotomous or coarsened representations of migration flows, we use recent innovations in valued exponential-family random graph modeling (Valued ERGMs or VERGMs) to estimate a fully quantitative model of interdependent U.S. migration flows at the county level. Motivated by the popular discourse surrounding the California Exodus and existing theoretical and empirical research regarding U.S. internal migration, we focus on four potential social forces that contribute to population redistribution. They include costs of living, political environments, levels of urbanization, and racial demographics.

This relational view offers new opportunities for insight, but also poses challenges. For instance, interpretation of the relationship between nodal or dyadic attributes' impacts on migration (i.e., covariate effects) can be complex, as such relationships are subject to both the origin's and the destination's attribute values, and they can take various functional forms. Further, the superposition of forms from multiple effects can make the model difficult to interpret. Such complexities reflect the inherent challenges of capturing an interactive system in quantitative detail, and are thus not unique to migration systems, but are particularly acute when considering networks with valued edges. We here propose a visualization protocol that showcases how multiple mechanisms involving origin and destination attributes combine to influence the expected number of migrants between origin and destination regions. We utilize this approach to display how the political, racial, rurality, and housing covariates influence the predicted migration flow intensity across different scenarios, offering a quantitative exploration the impact of dyadic factors on migration.

Another advantage of the VERGM approach is that it offers *generative models,* which can themselves be used to probe the effects of inferred or hypothetical mechanisms beyond the dyadic level. Here, we use our empirically-calibrated migration model to perform *in-silico knockout experiments* to investigate how various social, economic, and demographic mechanisms contribute to observed patterns of population redistribution - including, specifically, maintenance of the California Exodus. These knockout experiments simulate migration flow networks under counterfactual scenarios where certain social effects are inoperative (Huang and Butts, 2022). Comparing the extent of California's relative net migration loss in the knockout scenarios with that in the observed scenario offers quantitative insights about the impacts of social effects on the pattern of population redistribution.

The remainder of the paper proceeds as follows. We begin in Section 2 with a brief review of different approaches to modeling migration systems, and the extant empirical research that motivates our hypotheses regarding population redistribution in the U.S. Section 3 describes the data and variables we use, the model setup including the functional form specification, derivation of the visualization protocol, and the knockout experiment procedure. In Section 4, we first offer an overview of the population redistribution pattern in the United States, and the pattern of net migration exchange between U.S. states. We then report our findings regarding the drivers of migration patterns from the ERGM analysis, and show how contributing effects can be visualized. The section concludes with results from knockout experiments. The last section summarizes our empirical findings, our contributions to the mathematical modeling of complex social systems, and some directions for future work.

2 Background

2.1 Modeling Migration Systems

Migration flows among geographical areas form a complex system, a perspective that has received extensive theoretical discussion in migration studies, in the school of *Migration Systems Theory* (MST, Bakewell et al., 2016; DeWaard and Ha, 2019; Fawcett, 1989; Kritz et al., 1992; Mabogunje, 1970). MST introduces two insights regarding migration. First, a migration system consists of flows of people, goods, information, cultures, and other institutions that interact with each other (Bakewell, 2014). This suggests that understanding migration processes demands a comprehensive survey of various factors and mechanisms, incorporating economic, political, geographical, and demographic analyses. Second, MST emphasizes the *interdependent* feature of migration systems, reflected in their conceptualization of "internal dynamics" (de Haas, 2010) or "feedback mechanisms" (Bakewell, 2014). The central idea is that there exist endogenous processes, where change in one part of the system can diffuse and alter other parts, creating a systemic momentum. This means that migration flows are correlated to each other. For instance, the migration flow from Seattle to Chicago is associated with the reverse flow from Chicago to Seattle, partly because migrants can carry social connections and useful information from their origin to their destination, motivating and facilitating migration in the reverse direction. Such interdependence among migration flows requires mathematical models of migration to account for the autocorrelation among their observations, and ideally, to also formally and explicitly describe the structure of the dependence.

Researchers have developed various methods to model migration across disciplines including econometrics, geography, statistics, and sociology. A convenient and widely used approach is to treat migration as a feature of areal units, analyzing how the characteristics of a place are associated with marginal migration rates into and out of it (e.g., Partridge et al., 2012; Treyz et al., 1993). This approach has offered many useful insights and serves as a powerful framework for building predictive models of demographic change (Azose and Raftery, 2015, 2018). Methodologically, techniques to account for the autocorrelation in this data structure (areal/lattice data) are well developed in spatial statistics (Banerjee et al., 2014). However, migration is by nature a *relational* process between two places: origin and destination. The above approach by construction marginalizes migration either from an origin perspective or a destination perspective (or condenses both), obscuring how origin and destinations jointly and interactively shape the migration flows between them; such interactions are known to be of considerable importance, as articulated in the classical "push-pull" factor model (Lee, 1966) of migration. From a network analytic perspective, such models are equivalent to modeling the migration network purely in terms of expected outdegree and indegree effects (sometimes called *expansiveness* and *popularity* in the ERGM literature (Holland and Leinhardt, 1981)). Although simple, such models are very constraining - they are essentially similar to a single-dimensional singular value decomposition (SVD) approximation of the adjacency matrix - and are limited in their ability to represent complex structure.

A second model family is the so-called "gravity model" (widely used in spatial econometrics), whose unit of analysis is no longer a geographical area but flow within an ordered pair of geographical areas (i.e., an *edge variable*). The original idea of this model family is that the extent of migration flow from origin i to destination j (M_{ij}) is positively associated with population sizes in origin and destination (P_i, P_j) and negatively associated with the distance between (D_{ij}), with the decay usually posited to follow a power law (Zipf, 1946), thus superficially resembling gravitational attraction.[1] Formally, this family is written as

$$M_{ij} \approx C \cdot \frac{P_i^{\alpha} \cdot P_j^{\beta}}{D_{ij}^{\gamma}},$$

where C, α, β, γ are positive parameters. Although nonlinear on its original scale, the power law model is intrinsically linear, as shown via the log space representation

$$\log M_{ij} = \mu + \alpha \log(P_i) + \beta \log(P_j) - \gamma \log(D_{ij}) + \varepsilon_{ij}.$$

where $\mu = \log C$ and log error ε_{ij} are unknowns. Factors other than distance and population size may be incorporated by choosing a suitable regression form for μ. The linear form has facilitated further elaboration, e.g. using a GLM structure to capture discrete outcomes (e.g., Biagi et al., 2011). Although the gravity model does not provide a means of specifying dependence among flows, some extensions in this direction have been proposed (see reviews by Patuelli, 2016; Poot et al., 2016).

The gravity models have always been in close relationship with network models, with abundant shared knowledge and mutual development. Fundamentally, gravity models constitute a particular class of network regression models (albeit not necessarily OLS network regression, e.g. Krackhardt (1988)), a very flexible and successful family. Substantively, the functional form of the gravity model arises naturally as a model for *tie* (or interaction) *volumes* between regions under power-law spatial interaction functions, a widely observed functional form for interaction probabilities at the individual level (Butts and Acton, 2011); this, along with the strongly predictive power of distance itself for social networks (Butts, 2003), has been argued to provide a mechanistic explanation for why aggregate interactions are often well-approximated by gravity models (Almquist and Butts, 2015). The identification of gravity models with network regression also points to their limitations: while very flexible in specifying relationships between covariates and tie values, network regression models do not specify dependence among edge variables. While workarounds such as quadratic assignment procedure (QAP) tests (Dekker et al., 2007; Krackhardt, 1988) can provide statistical answers that are robust to dependence effects, parameterization and/or generation of networks with dependence requires other approaches.

The specification of models for networks with complex dependence among edge variables is a major concern of work on exponential-family random graph models, which we discuss in detail in Section 3.2. ERGMs provide a rich language for specifying interdependencies among edges, as well as associated statistical theory and methodology for inferring such dependencies from observed network data. Importantly, ERGMs are *generative* - i.e., they provide a full probability model for the target network, and thus can be used for hypothetical realizations of an inferred data generating process. This makes them especially well-suited to mechanistic investigation using approaches such as *in silico "knockout" experiments* and other computational techniques. The increasing availability of scalable and valued-data ERGMs opens the door to modeling migration systems in a substantively-richer and more statistically-rigorous way.

As noted, one advantage that ERGMs have is the ability to explicitly and formally describe the interdependence of edges within networks. In connection with MST, researchers have utilized this feature to formalize and test the patterns and mechanisms of the endogenous feedback processes in migration systems

[1] This formulation is also used to describe other types of spatial interactions such as international trade; see e.g. the review of Anderson (2011).

(Huang and Butts, 2022; Leal, 2021; Windzio et al., 2019). Specifying dependence structure can also improve statistical inference. The autocorrelation among migration flows can not only introduce associations in residuals, but may as well impose more general autoregressive structure. In this case, methods that focus on correcting for correlation in the residuals (e.g., QAP) could be insufficient, running the risk of failing to account for the impact of endogenous factors on covariate effects.

Likewise, the generative aspects of ERGMs are particularly relevant in the context of studying migration systems. The ability to simulate from empirically calibrated or *a priori* models allows researchers to extrapolate models across spatial and temporal contexts and even investigate counterfactual scenarios. Although there is work in this direction (including applications to the study of migration systems (Huang and Butts, 2022)), it is arguably an under-appreciated property of this model family, which has been mostly employed as a tool for hypothesis testing. This paper aims to exploit the generative capacity of ERGMs to quantify the contribution of various drivers of population redistribution to the California Exodus.

Despite these advantages, using ERGMs to study migration systems poses a number of challenges. First, it can be computationally intensive to fit (and sometimes to simulate draws from) such models, since closed-form (or even directly computable) expressions for the likelihood are not attainable except in special circumstances. Moreover, generative models for valued/weighted networks are less developed than binary networks, in terms of formal specifications of dependence structures, theoretical justifications of those specifications, and efficient computational tools; this means that researchers sometimes have to dichotomize migration flows, losing critical information about the scale of migration flows. While it is not the focus of the paper to advanced generative models for valued/weighted networks, we employ recent advances in this area to offer a quantitative understanding of the population redistribution pattern within the United States.

Moving beyond ERGMs *per se*, a general challenge in modeling relational data such as migration system data is understanding the combined effects of multiple influences, since prediction of a specific migration flow usually involves attributes from different sources (e.g., origin and destination) that can be combined in different ways. The usual approach of interpreting coefficients separately under the *ceteris paribus* condition is often unhelpful here, as these covariates are intrinsically inter-related. For example, often it is substantively natural to include covariate factors (e.g., housing costs) of origin, destination, and their absolute difference, where the last term can no longer be interpreted only as a dissimilarity measure since the statistic is fixed once we hold constant the origin and destination covariates. This paper tackles this problem by introducing a visualization protocol that helps interpret the multiplex of inter-correlated functional forms that is common in relational data analysis.

2.2 Drivers of Population Redistribution

This section examines possible drivers of population redistribution, with an empirical focus on the case of California Exodus. The first potential driver is the cost of living, suggested by the allegation that people migrate out of California because it is too expensive to live in (e.g., Bahnsen, 2021; Beam, 2021). This is in correspondence to the neoclassical economic theory of migration, that migration happens when the move brings net profit, and lower living costs in destination can be a substantial source of net profit. This motivates our hypothesis:

H1: The migration rate from origins with high costs of living to destinations with low costs of living is higher than the reverse.

Following the popular narrative that the California Exodus is a political outcome (Bahnsen, 2021), we hypothesize that political environment could also serve as a driver of population redistribution. Public choice theory and the consumer-voter model consider migration as a means of realizing people's policy preferences (Dye, 1990; Tiebout, 1956). Empirical research on U.S. internal migration has also repeatedly observed Americans "voting with their feet" (Huang and Butts, 2022; Liu et al., 2019; Preuhs, 1999; Tam Cho et al., 2013). The allegation that Californians leaving their liberal state behind are "leftugees" fleeing Democratic

governance (Dorsey, 2021) motivates our second hypothesis:

H2: The migration rate from liberal-leaning origins (i.e. those with higher share of supporters for the Democratic Party) towards conservative-leaning destinations is higher than the reverse.

Since population redistribution goes hand in hand with urbanization (Lichter and Brown, 2011; Ravenstein, 1885), it is possible that California Exodus is a reflection of the ex-urbanization process. Henrie and Plane (2008) and Plane et al. (2005) documented the shift of U.S. population from urban areas to rural areas in the 1990s. If this is still happening in 2010s, that might be an underlying mechanism behind California's net migrant loss. We therefore hypothesize that:

H3: The migration rate from urban origins to rural destinations is higher than the reverse.

Last but not the least, racial dynamics play a critical role in American lives, including migration decisions (Crowder et al., 2006, 2012). According to the literature, "White flight" is a frequently observed phenomenon (Boustan et al., 2023; Frey, 1979; Woldoff, 2011), where members of the non-Hispanic White population migrate out of racially-diverse places and settle in White-dominant areas. While White flight is associated with the ex-urbanization process, previous literature has identified racial factors to be a unique and non-negligible contributor to this movement (Frey, 1979; Kruse, 2013). Considering California's diverse racial demographics, White flight could hypothetically contribute to the exodus, and we thus hypothesize that:

H4: The migration rate from origins with low non-Hispanic White concentration to destinations with high non-Hispanic White concentration is higher than the reverse.

These hypotheses embody a combination of conventional wisdom and notions motivated by migration patterns seen elsewhere. But are any of them true - and, more importantly, can they account for the California Exodus? For this, we turn to our empirical analysis.

3 Materials and Methods

3.1 Data

We model the inter-county migration flow network among all 3,142 U.S. counties. The outcome of interest is the average number of migrants moving between each directed pair of counties each year during 2011-2015, which is calculated and released by the American Community Survey (ACS) administered by the U.S. Census Bureau.

The key covariates capture the characteristics of origin and destination in their costs of living, political climates, level of urbanization, and racial compositions. The cost of living is measured by the median housing costs in 2006-2010 ACS; the political climate is represented by the percentage of voters that voted for the Democratic candidate (Obama) in the 2008 presidential election, as that was the latest national-level election before the study period. The level of urbanization is indicated by the proportion of rural population of a county, estimated by the 2010 Decennial Census. Lastly, the feature of a county's racial composition is described by its Non-Hispanic White population in the 2010 Census, as this is the most populous racial-ethnic category in the U.S.

The model also considers other covariates that can potentially influence the magnitude of migration flows. The demographic covariates include the (log) population size, log population density (in thousand people per squared kilometers), and age structure (potential support ratio, PSR: ratio of population that are 15-64 years old over population that are 65+ years old), all using 2010 Census Data. The economic covariates include percentage of renters (in contrast to home owners) using 2010 Census, unemployment rates, and percentage of population with higher education attainment, both using 2006-2010 ACS. The geographic covariates include the log distance between origin and destination counties (in kilometers), a dummy variable indicating whether they belong to the same state, and fixed effects for the four major U.S. regions (Northeast, South, Middle West, and West). We also include log migration flow in the previous time

period (2006-2010) of the focal migration flow, and the network dependence terms specified in the following section.

3.2 Valued ERGMs

We first model the migration patterns using the valued exponential-family random graph models (valued ERGMs, or VERGMs) (Krivitsky, 2012). The ERGM is a parameteric generative model that impose an exponential family distribution to describe the network structure of interest:

$$\Pr(Y = y|\theta, X) = \frac{h(y)\exp(\theta^T g(y,X))}{\sum_{y' \in \mathscr{Y}} h(y')\exp(\theta^T g(y',X))}, \tag{1}$$

where Y is the random variable of network with realization y. $g(\cdot)$ is a vector of sufficient statistics with corresponding parameters θ. The sufficient statistics can be flexibly specified to incorporate both structural covariate effects (e.g., housing price differences) and endogenous dependence terms that capture autocorrelations among migration flows. In this paper, we include two dependence terms, mutuality and waypoint flow, to account for the endogenous mechanisms that contribute to the symmetry at the dyad-pair level and the node level, beyond the specified covariate effects. Mutuality captures the scale of *reciprocated flow* within dyad pairs ($i \to j, j \to i$) by calculating the summation of the minimum edge value across all dyad pairs:

$$g_m(y) = \sum_{(i,j)\in\mathbb{Y}} \min(y_{ij}, y_{ji}). \tag{2}$$

The larger the reciprocated flow within a dyad pair, the larger the statistic. For example, if there are 6 migrant exchange between counties i, j, a distribution of $\{3,3\}$ will have the largest reciprocated flow and the corresponding statistic (3), and a distribution $\{0,6\}$ will have the smallest (0). Therefore, a positive coefficient will indicate an endogenous pattern of dyad-level reciprocity, and vice versa. The waypoint flow takes a similar formula, but captures the volumetric flow through each node by examining its total inflows and outflows:

$$g_f = \sum_{i\in\mathbb{V}} \min\{\sum_{j\in\mathbb{V}, j\neq i} y_{ij}, \sum_{k\in\mathbb{V}, k\neq i} y_{ki}\}. \tag{3}$$

The larger the volumetric flow moving in and out of a node, the larger the statistic. A positive coefficient will indicate an endogenous pattern of node-level symmetry, and vise versa.

$h(y)$ is a reference measure that determines the probability distribution of the networks when $\theta \to 0$. As a VERGM, since our outcome of interest is the count of migrants between two counties, we specify the shape function as a Poissonian reference measure:

$$h(y) = \prod_{(i,j)\in\mathbb{Y}} (y_{ij}!)^{-1} \tag{4}$$

This amounts to the assumption that migration events are indistinguishable within edges. The denominator of the equation 1 is the normalizing factor that defined on $\mathscr{Y}$, the set of all possible network configurations based on the same vertex set. This intractable function is the source of computational complexity for ERGMs, as it is a function of both the parameter to be estimated, and the set of possible network structures. This is especially the case for VERGMs, since each dyad now can take not only two values for binary networks, but all natural numbers. The more than three-thousand nodes also increases the computational load of our model. To grapple with this challenge, we employ a parallelizable Maximum Pseudo-likelihood Estimation procedure for VERGMs (Huang and Butts), which is efficient and shows good estimation quality for high-edge-variance networks such as ours.

3.3 Functional Form Specification

There are many possible functional forms for network models even just considering linear formats, since the edge-based models jointly account for the covariates of origin and destination. We thus formulate our key covariate effects based on our theoretical assumptions of their mechanisms that influence migration.

For the cost of living, we include the housing costs of origin and their the difference between destination and origin (destination minus origin). Drawing on the aspiration-ability model of migration (Carling, 2002; Carling and Schewel, 2018), we posit that origin housing costs influences people's financial well-being, which translates into their capacity to migrate; the difference in housing costs influence the utility gain of migrating, altering their aspiration of the migration.

In terms of political, rurality, and racial covariates, we include a dissimilarity measure, implemented as the absolute difference between origin and destination in the corresponding covariate. This follows the operationalization of previous literature (Huang and Butts, 2022), which reveals a segmental effect in which less migration happens between counties with larger difference in political climates, levels of urbanization, and racial compositions. Since our interest is population redistribution generated from asymmetric migration, we further include two directional effects. The first is the covariate level of the origin, and the second is a sign function (+1 when destination has a higher covariate level than origin, -1 when the reverse, and 0 when equal).

3.4 Visualizing Functional Forms

The composite functional forms of each covariate effect pose the question of how to unpack and interpret their joint effects. We develop a visualization protocol that tackles this problem. For each functional form, the protocol calculates the expected edge value under each possible combination of the covariate value of the origin and destination. To make it more comparable across functional forms, we then normalize it by calculating the ratio of this expected value over the expected value that would be obtained if both origin and destination took the average observed value of the covariate.[2] We describe this formula as follows.

In the absence of dependence terms, a Poissonian VERGM is identical to a network regression model with a independent Poisson distributions on each edge (Krivitsky, 2012), where there expected value of the i, j edge is:

$$\mathbb{E}(Y_{ij}) = \exp(\theta^T \Delta_{ij}^{0\to 1} g(y, X)) \tag{5}$$

where Δ denotes the change in the sufficient statistics when the focal edge's value goes from zero to one. If we only focus on one covariate $f_k(X_{ij})$ (whose sufficient statistic in ERGM will be $f_k(X_{ij}) \cdot y_{ij}$), then we have:

$$\mathbb{E}(Y_{ij}) \propto \exp(\theta_k \cdot f_k(X_{ij})) \tag{6}$$

so we can express the conditional expected value as a function of origin's and destination's covariate level by calculating the exponentiated product of the functional form and the corresponding coefficient in equation 6. We further add a normalizer to center the expected value and make it more comparable across different functional forms. The normalizer is the expected edge value when the covariate of the origin and destination is set to the average value (described in the previous footnote) across the vertex set (X_0):

$$\mathbb{E}(Y_{ij}) \propto \exp(\theta_k \cdot [f_k(X_{ij}) - f_k(X_0)]) \tag{7}$$

The formula is in essence the ratio between the expected value of a focal edge under a specific origin-destination covariate vector over the expected value where the origin and destination has the covariate value equal to the average value.

[2]For political, racial and rurality covariates, we use the population-weighted national mean, treating every county as if it had the same share of Democratic voters, non-Hispanic Whites and rural population as the national percentage. For housing prices, we use the national median, as the functional form takes the logarithm of the prices.

When we need to calculate the ratio for composite expected value, we can simply take the product of their ratios for each form. In the Results section, we will display the functional form of both separate effects (e.g. origin housing costs) and composite effects (e.g. origin housing costs plus difference in housing costs).

Note that this is not exactly the same as the conditional expectation ratio in our specified model, since the model contains dependence terms that distort the edge distribution away from a regular Poisson distribution. A rigorous calculation of the exact expectation ratio is, however, computationally prohibitive, as it requires numerical integration of all possible edge values times their probability function for every realization of the covariate vector. Nevertheless, the knockout experiment in the following subsection takes the dependence into control, offering a closer look at the functioning of the VERGM with dependence terms.

3.5 Knockout Experiments via Network Simulation

The visualization of functional forms offers structurally "local" insights about how each social force influences migration patterns. Building upon that, we want to quantify how theses social forces contribute to the social phenomenon of interest on a global scale, specifically population redistribution and the California Exodus. We achieve this by leveraging the generative feature of ERGMs to perform *in silico* knockout experiments via network simulation. A *knockout experiment* as employed in a social science context is a model-based thought experiment that examines counterfactual scenarios where certain posited social forces are inoperative, while all other forces are left at their observed levels (Huang and Butts, 2022). The change in outcomes of interest relative to the behavior of the full model is used to probe the impact of the knocked-out mechanism. Here, we implement knockout of mixing effects by simulating migration flows with all counties having their covariates of interest fixed at an identical value average that is specified in the previous footnote (removing differential mixing). Simulating flows obtained under these conditions, we compare California's ranking in net migration loss across all states under the knockout scenarios with the observed models. This allows us to probe the connection between the mechanisms captured by the model and our social phenomenon of interest. For example, if under the hypothetical condition where every U.S. county has the same housing cost, California's relative net migration loss is not as severe as the observed situation, it would suggest that housing-cost effects on migration could be a contributor to the California Exodus; by turns, if eliminating housing disparities has no impact on asymmetric migration, we can rule it out as a driver of migration loss.

To assist the interpretation of the quantitative results from knockout experiments, we include positive and negative controls in simulation, alongside knockouts of our key covariates of interest: political, racial, rurality, and housing attributes. Originating in the experimental sciences, positive and negative controls are experimental conditions that researchers expect to produce positive and null results, respectively; the controls validate the experimental procedures, serving as the benchmark for other regular experimental settings. In an *in silico* setting, controls remain important to verify that the model is sensitive to manipulations that should have an impact on the outcome of interest (and, by turns, that it is not overly sensitive to manipulations that should not have an impact). Here, we knock out distance effects as a negative control, treating all dyads as having a common log distance set at national mean. We expect the knockout of non-directional distance effects to not alter the rankings of net migration loss across the country, and the difference between this case and the full model can be considered as a combination of numerical noises and some second-order impacts (since we include complex network dependence terms). The removal of population effects by equally distributing population across all counties serves as a positive control case, as we expect the removal of population effect to have a large impact on the population redistribution pattern. The purpose of these two controls is not substantive interpretation of the fundamental distance and population effects, as the counterfactual scenario is arguably radical and unrealistic, but rather, to provide insights into the question of "how small is small" and "how big is big" in terms of altering migration ranking.

4 Results

4.1 General Patterns of Population Redistribution

Table 1: Annual Population Change in the United States, 2011-2015

	Count	Crude Rate (%)
Population	308,739,316	
Natural Change		
Births	3,961,037	1.28
Deaths	2,598,956	0.84
Natural Increase	1,362,081	0.44
International Migration		
Immigration	1,841,695	0.60
Inter-county Migration		
Total migrants	17,176,675	5.56
Node-level asymmetry	1,523,550	0.49
Dyad-level asymmetry	3,844,434	1.25

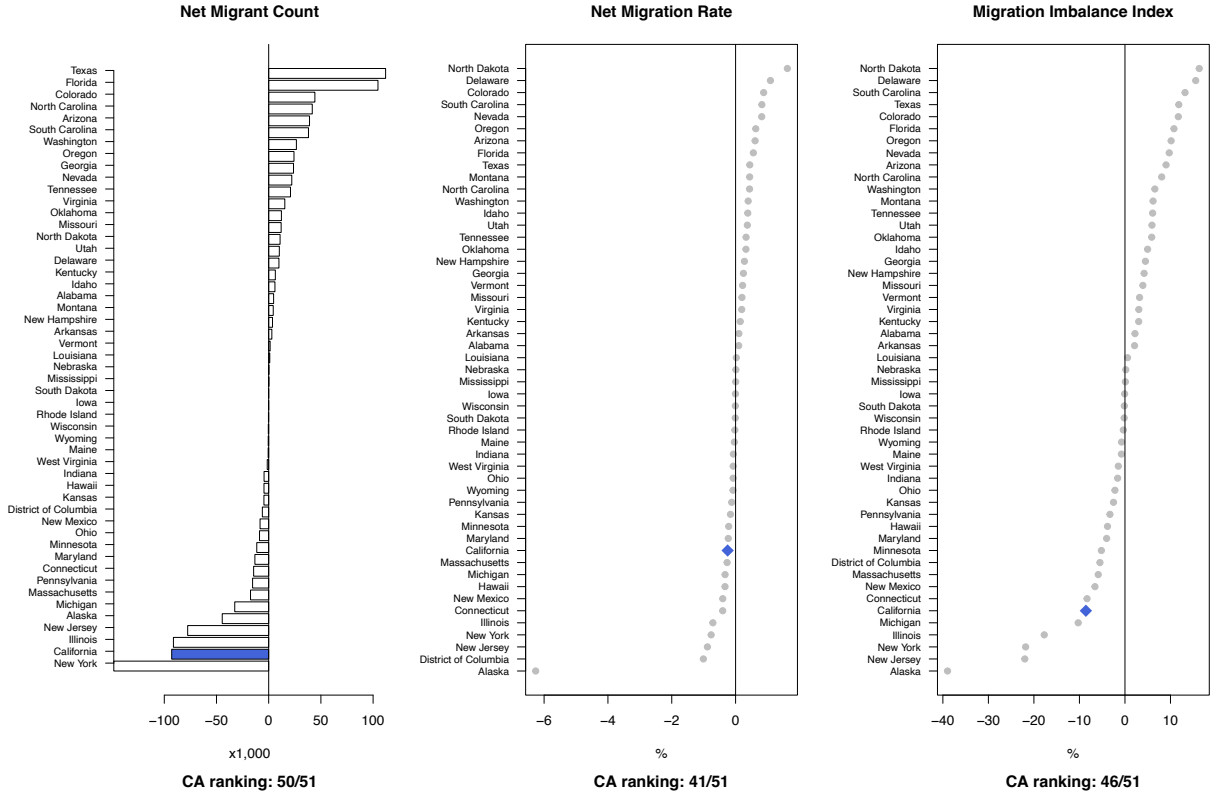


Figure 1: Net migrant count, net migration rate, and migration imbalance index by state

To offer a broad view of population change in the study period, Table 1 shows the annual population changes from different demographic processes and their crude rates (normalized by the total population

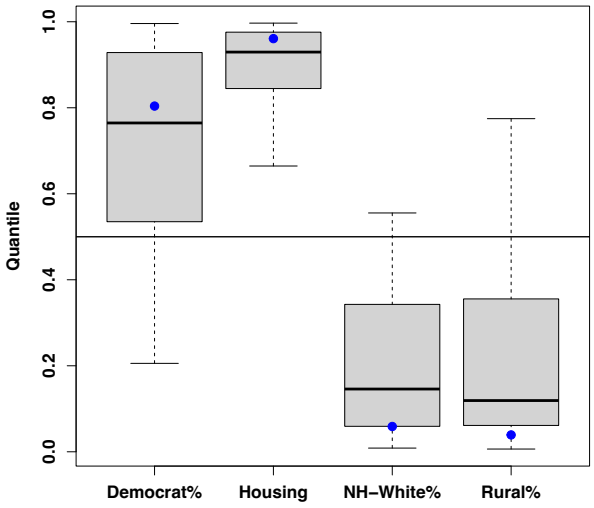


Figure 2: Quantiles of attributes for California counties (boxes) and the state as a whole (blue dots) relative to all U.S. counties

size).[3] Compared to natural change and international migration, inter-county migration in the U.S. is a more substantial demographic process with a larger share of population involved. When it comes to population change, the asymmetric internal migration is similar to the scale of immigration and natural increase, all of which have a modest share of population, whcih is around 0.5% to 1%. This confirms that as a developed country, the U.S. has a relatively modest population change in the 2010s (Rees et al., 2017).

Figure 1 examines the phenomenon of the California Exodus by comparing the net migrant loss of California (shaded in blue) across three metrics against other U.S. states and the District of Columbia (DC). The left panel displays the net migrant count, which is the total in-migrants minus the total out-migrants. It shows that California has a large net migrant loss, only second to New York among the 51 states and DC.

Yet, considering the fact that California is the most populous state (roughly 25% more than the second populous state, Texas, in 2010), the middle panel calculates the net migration rate, which is the net migrant count divided by the state's population. The normalized metric observes California to have a less extreme net migration loss. While it still ranks at the lower end of the list, it is not very different from the majority of the U.S. states, which are within the range of -1% to 1%. In other words, the large net outflows of migrants from California can be partly explained by its largest population size.

Although the middle panel may suggest that there is nothing to be explained - the California Exodus is simply a size effect - examining the *relative asymmetry* of migration to and from California gives a richer picture. The right panel calculates the migration imbalance index (MII) of each state, which is the net migrant count divided by the sum of in-migrants and out-migrants.[4] The measurement indicates the

[3]The population size comes from 2010 Census, the natural change data comes from U.S. Center for Diease Control and Prevention, and the international and internal migration comes from ACS 2011-2015. The natural increase is the number of births minus the number of deaths. The dyad-level asymmetry is the sum of absolute difference across all dyad pairs divided by two: $A_d = \frac{\sum_{i,j}|Y_{ij}-Y_{ji}|}{2}$, and the node-level asymmetry is the sum of absolute difference across all nodes in their inflows and outflows divided by two: $A_n = \frac{\sum_i|\sum_i Y_{ij}-\sum_i Y_{ji}|}{2}$.

[4]MII coincides with the migration efficiency/effectiveness index in some migration literature (Bell et al., 2002; Shryock et al., 1973). It is also directly related to the external-internal (E-I) Index in social network analysis (Krackhardt and Stern, 1988), although

proportion of related migrant flows that are inflows of a focal place, capturing the level of imbalance between inflows and outflows of migrants. The right panel reveals that migration imbalance generally has larger variation across states than the net migration rate, as the former focuses on a smaller population, i.e. the migrant population. California has relatively lower ranking in migration imbalance than net migration rate, and its value is farther away from other U.S. states, suggesting a noticeable imbalance in its in/out-migration flows.

In summary, Figure 1 reveals that California is indeed experiencing net migration loss, although the severity relative to other parts of the country vary by the metric we read. Moreover, despite the popularity of the California Exodus narrative, California is actually not the place with the most net migration loss: the New York state has stronger net loss than California across all metrics, and the net migration rate and migration imbalance of Alaska is substantially lower than the rest of the states. These other cases poses important empirical questions that future research should consider.

Lastly, as we consider the possible contributor of California's outstanding net migration loss, we examine California's attributes in Figure 2. The boxplots shows the quantiles of California counties in those attributes across all U.S. countires, and the blue dots indicates the quantile of California across the 51 states and DC. Compared to other parts of the country, California is indeed a place with stronger left-leaning political environments, expensive housing, larger racial and ethnic minority population share, and higher levels of urbanization. These dimensions are characteristics where California stands out, and therefore has the potential of explaining its migration patterns.

4.2 Functional Forms of Migration Driving Forces

4.2.1 Estimated Effects

Table 2: Valued ERGM for Inter-County Migration Flows, 2011-2015

	Estimate	Std Err
Political Covariates		
Dissimilarity P(Democrat)	-.257***	.007
Origin P(Democrat)	.024**	.009
To higher P(Democrat)	-.008***	.001
Racial Covariates		
Dissimilarity P(NH-White)	-.172***	.006
Origin P(NH-White)	-.044***	.007
To higher P(NH-White)	.011***	.001
Rurality Covariates		
Dissimilarity P(rural)	-.457***	.004
Origin P(rural)	.330***	.006
To higher P(rural)	.018***	.001
Housing Covariates		
Origin log(costs)	-.283***	.005
Difference log(costs)	-.148***	.004
Control Covariates	(included)	

Note: $^{**}p < 0.01$; $^{***}p < 0.001$ (two-tailed tests).

To explain the underlying patterns of intercounty migration, we estimate a VERGM for the migration

the latter focus on external flows, so MII is equal to one minus the E-I index.

flow network, with the results of the key covariates of interest listed in Table 2. The model suggests that, on average, less migration happens between counties with larger differences in their political climates, rurality, and racial compositions as reflected by the percentage of the non-Hispanic White population. In terms of directional effects, the model predicts larger migration flows from counties with higher Democratic Party voter share, and towards counties where the Democratic party voter share is lower. The directionality of the political effects is largely in correspondence to the "lefugee" Hypothesis 2 that population are generally leaving from Democratic-party-leaning areas towards Republican-party-leaning areas. The racial effects also run in the direction predicted by the "White flight" Hypothesis 4. Holding other factors constant, counties with smaller proportions of non-Hispanic White population send out more migrants, and larger migration flows exist along the way that lead to a county with a higher share of non-Hispanic White population.

When it comes to rurality, the model is consistent with the ex-urbanization Hypothesis 3 that migration flows are larger when they are moving towards counties with a higher share of rural population than the origin. Yet, the model also shows that counties with higher rurality on average send more migrants out than those with lower rurality. In other words, more migration flows are moving towards more rural regions, but more of them come from a rural county. The housing effects also offer mixed evidence in light of the neoclassical-economic Hypothesis 1. Although migration flows are larger where moving brings greater declines in housing costs from origin to destination, counties with lower housing costs also observe larger out-migration flows. This means that migration typically happens from places with inexpensive housing to places with even less expensive housing.

The model also controls for a series of other covariate effects and endogenous dependence structure, reported in Table A1 in the Appendix. The positive mutuality and the negative waypoint flow patterns suggest that, holding other covariate effects constant, the observed migration flow network is more reciprocal at the dyad-pair level and less symmetric at the node level than a random network configuration. This implies the existence of endogenous network patterns discussed in prior literature (Leal, 2021; Huang and Butts, 2022). For example, the practice of return migration could promote dyad-level reciprocity, and the signaling effects of county attractiveness can lead to endogenous node-level asymmetry (large migration inflows signaling the popularity of this county, retaining potential migrants from leaving, resulting in an imbalanced in&out-flow of the county).

4.2.2 Visualizing Functional Forms

For a typical research paper using parametric models, the results section usually stops at the previous subsection, after summarizing whether the directionality of the key effects confirms or refutes the hypothesis. While it is informative to use parametric models as tools for hypothesis testing by evaluating their qualitative behavior, there are more insights one could gain from further the examination of the models.

First of all, besides the signs of the coefficients and their corresponding p-values, their magnitudes also carry critical information about the scale of the effects of interest. Taking the political covariates in Table 2 as an example, the coefficient of origin effects and binary directional effects look an order of magnitude smaller than that of the dissimilarity effect. However, it is difficult to directly interpret the parameter magnitudes, which is subject to the scaling of the covariate distribution.

The second question is about how to interpret holistically the effects of interest, as the different effects (origin, difference, dissimilarity) are interdependent, and holding other factors constant to interpret each single functional form can be unrealistic. This could be a critical question as sometimes different effects offer mixed evidence about substantive hypotheses, such as the rurality and the housing effects in our model. It is of substantive interest to understand how these different effects jointly shape the migration pattern.

To quantify the magnitude of the modeled effects and more concretely understand the separate and joint roles of the functional forms, we visualize the (normalized) predicted migration flow size as a function of origin's and destination's covariate values, displayed in Figure 3. Each row presents one chunk of covariate

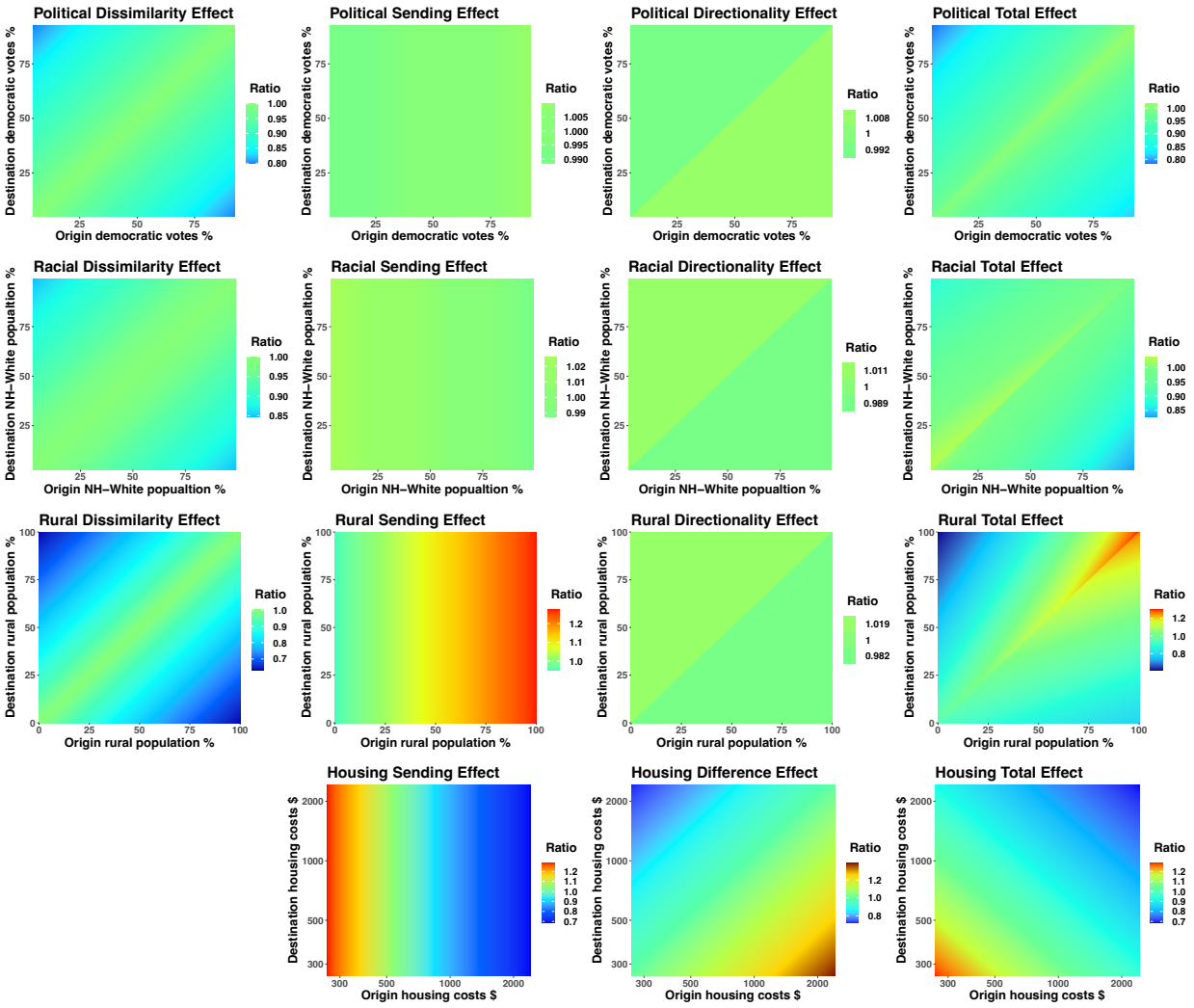


Figure 3: Function forms for political, rural, racial and housing effects

effects, and each column presents a type of functional form, where the higher value in the heatmap indicates the model predicts the migration flow to be higher under these origin-destination covariate values.

The first row of Figure 3 shows that the directional functional forms (sending and directionality effects) produce very little alternation of the expected migration flow, compared to the undirectional functional forms (dissimilarity effects). The middle two panels show a tiny gradient in its coloring, and the total effects largely resemble the dissimilarity effect, suggesting that the sending and directionality effects make little contribution to the overall effect of political climate. Similarly, in the second row, directional effects of racial covariates also appear negligible, and the undirectional dissimilarity effect dominates the total effect of racial composition. These visualizations tell us that while the directional effects of political and racial covariates run in the direction that correspond to the hypotheses, their effect sizes are small compared to the nondirectional dissimilarity effects.

In the third row of Figure 3, although the directionality effect of rurality still resembles those of the previous political and racial covariates, bringing small variation in the expected migration scale, the rural sending effect is strong, and alters the rural total effect to be asymmetric. The bottom row shows that while the sending and difference effects predict substantial variation of expected value across different housing

values, their combination offsets each other in the bottom right panel; the gradient of the total effect largely evolves along the $y = x$ line, meaning that swapping the housing costs of origin and destination does not leads to major change in the expected migrant counts. This means that the total housing effect is largely symmetric.

4.2.3 Visualizing Functional Forms: The San Francisco County Case

To further aid our interpretation of the total effects, Figure 4 examines the case of San Francisco (SF) county, California, and evaluates its expected migration flows towards and from other counties based on their corresponding covariate value. The first column is a replication of the last column in the previous figure, but adds reference lines that indicate the covariate level of SF county. The middle column extracts from these two reference lines and plot the expected number of immigrants to (brown solid lines) and emigrants from (grey dotted lines) SF county as a function of the origin/destination county's covariate level. The upper right panel of each row summarizes the middle column by getting the difference of immigrant ratio and the emigrant ratio, where a positive ratio difference (shaded in solid brown lines) suggests an expected net migration gain for SF county, while a negative ratio difference (shaded in dotted grey lines) suggests an expected net migration loss for SF county. The bottom right panel of each row plots the histogram of U.S. population about the covariate level of their residing counties. The juxtaposition of the last two plots reflects whether the country's population gravitate towards counties that SF county has net migration gain from (shaded in brown), or counties that SF county has net migration loss towards (shaded in grey), offering a first-order approximation to whether the social effects promote or suppress population loss from a county like San Francisco.

Focusing on the right column of Figure 4, we observe that SF county receives net migration gains from counties with more Democratic-party voter share, which comprise a small share of U.S. population. By turns, it loses migrants to counties with less Democratic-party voter share, which comprise a large share of U.S. population. Similarly, in the second row, SF county receives net migration gains from counties with less non-Hispanic White population share, which comprise a small share of U.S. population. The functional form of rurality for SF county is a bit more complicated, as the county takes the extreme value of 0% rural population. The county is expected to have no net migration exchange with other counties that have 0% rural population, which consist 7% of the total U.S. population. SF county is expected to lose population to counties with rural population larger than zero but smaller than 13%, which includes about 51% of the total U.S. population. In other words, on average, there are slightly more persons residing in counties that SF county has net migration loss towards. However, once the county deviates from the extreme case of the fully urbanized, the trend reverses, with more of the U.S. population residing in places from which the focal county has net migration gain. Lastly, the bottom right panel shows that the majority of the U.S. population resides in counties with cheaper housing than SF county, areas to which SF would be expected (*ceteris paribus*) to lose population. Overall, for the SF county case, across all covariates, the model predicts an overall net migration loss from SF county; this is not because all factors *unilaterally* favor emigration from SF, but rather because in each case SF's attributes favor immigration from a relatively small number of counties (with relatively low total population) relative to those to which they favor emigration.

4.3 Knockout Experiments for the California Exodus

The visualization of covariate effects offers us some quantitative insights about how different social forces operate across different origin/destination pairs. However, our examination of the SF case underscores the intuition that the way in which such forces play out depends upon the global distribution of population (and covariates), which is challenging to infer from direct inspection. For instance, the high level of urbanization in SF county makes it an interesting but special case, and it becomes difficult to visualize every possible ru-

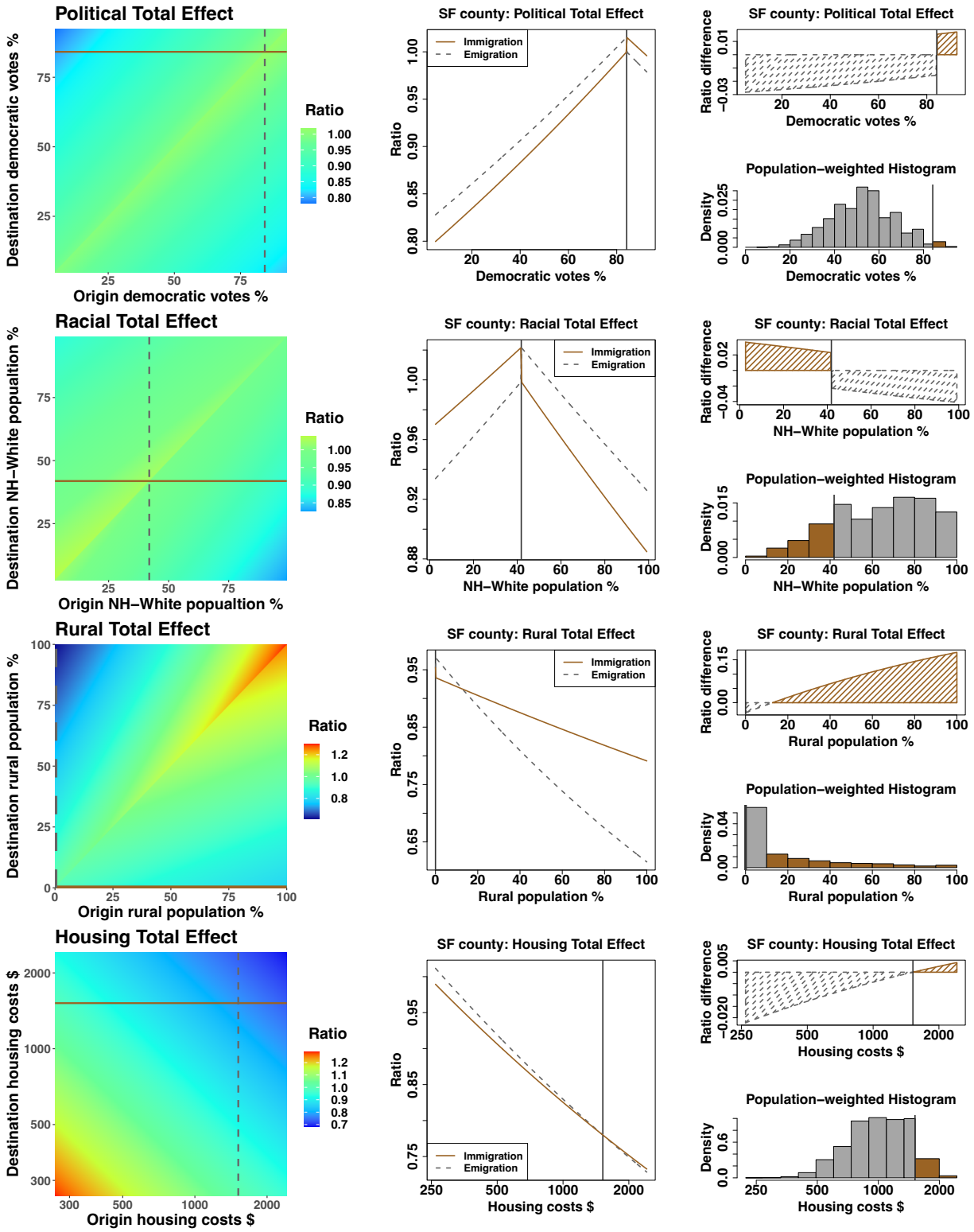


Figure 4: Function forms for migration effects involving San Francisco county. (left) Dyadic effects, with vertical and horizontal lines showing SF attributes. (center) Net immigration (solid lines) and emigration (dotted lines) effects for SF, given origin/destination county attributes; vertical line shows SF position. (right) Areas between curves (net immigration) from the center plot by origin/destination county attributes; histograms show population-weighted distributions of U.S. counties, with brown columns indicating population in net SF-immigration counties.

Table 3: California's Average Simulated Ranking with and without Knockouts, by Metric

	Net Migrant Count		Net Migration Rate		Migration Imbalance	
	Ranking	Change	Ranking	Change	Ranking	Change
Full Model	50.00		42.08		45.55	
Remove Distance Effect	50.00	0.00	41.92	-0.16	45.35	-0.20
Remove Population Effects	48.75	-1.25	28.30	-13.78	46.63	+1.08
Remove Political Effects	50.00	0.00	42.00	-0.08	45.16	-0.39
Remove Housing Effects	50.00	0.00	41.92	-0.16	44.95	-0.60
Remove Racial Effects	50.00	0.00	39.36	-2.72	42.49	-3.06
Remove Rurality Effects	50.00	0.00	43.00	+0.92	47.49	+1.94

rality level that California counties take and integrate them to offer a holistic evaluation of the rurality effect on the California Exodus. Building on these exploratory insights, this section aims to explicitly examine the connection between migration patterns incorporated into the model with specific social outcomes of interest, such as the California Exodus.

We achieve this by performing *in silico* knockout experiments, with results displayed in Table 3. The first column suggests that California's ranking in net migrant count stays constant throughout all the knockout scenarios except the positive control that knocks out population, contributing to a 1.25 position improvement its ranking (smaller ranking means less net migrant loss). Notice that only knocking out population effects in the positive control alters California's average ranking in net migrant count, and that in the second column, the net migration rates under normalized state population lead to fluctuations of California's average rankings under all knockout scenarios. This suggests that California's status as the largest U.S. state is a major explanation for its substantial net emigration in absolute terms.

In the second column of Table 3, the removal of political and housing effects improves California's ranking in net migration rate at a scale smaller than or roughly equal to the negative control of removing distance effects. Although political and housing effects seem to operate in a direction that contributes to California exodus as hypothesized, their influence on net migration rate is substantively negligible. Knocking out racial effects and rural effects improves and worsens California's relative net migration rate, respectively, indicating that racial effects contribute to California Exodus (from a migration rate angle), while rural effects actually buffer California from even larger population loss. These two changes are larger in their scale than the negative control of distance effects, but not comparable to the positive control of population effects, suggesting their impacts to be moderate.

The last column in Table 3 shows California's ranking of migration imbalance. As with the case of net migration rate, removing political, housing, and racial effects reduces California's relative migration imbalance, while removing rurality effects worsens it. Quantitatively speaking, the impact of knockouts of political and housing effects are again similar to that of the negative control of distance effect, while the removal of racial and rural effects bring a ranking change even larger than that from the positive control case of population effects. The small alteration from the positive case is understandable, as the origin and destination effects of population are not hugely different in our model (as well as in many other gravity models, Boyle et al. (2014)); while changing the total size of migrant population (symmetrically) can alter state rankings of net migration rate given a constant total population denominator, for migration imbalance that solely focuses on the migrant population, this is no longer the case. The fact that none of the knockouts alters California's relative migration imbalance in a sizable way suggests that California's migration imbalance does not result from one single social effect.

5 Discussion

Leveraging a large-scale valued network model, this paper studies population redistribution patterns in the United States, and in particular the heatedly discussed case of the "California Exodus." Our analyses show that California indeed experienced net migration loss in the 2010s, although its scale varies depending on the metrics one examines; the exodus is substantial in absolute terms but relatively small in its crude rate (count per capita), while still being fairly considerable in its imbalance between in-migration and out-migration flows. Valued ERGM analysis reveals the direction of the political, rural, racial, and housing effects on population redistribution, which largely work in directions that would contribute to net migration loss for highly populous counties like San Francisco. Knockout experiments further show that racial effects contribute to the California Exodus, rurality effects work *against* the California Exodus, and while political and housing effects contribute to the California Exodus, their effects are largely negligible. The scale of these effects on the California Exodus varies by the migration metric used, but none of the knockout scenarios (except a positive control case for population distribution) alter California's ranking in net migration loss in a substantial way. This suggests that the California Exodus is not governed by one single social effect, but is a joint outcome of complex systemic patterns.

Methodologically, this paper offers a roadmap that aids interpretation of composite functional forms in parametric relational models via visualization. It also demonstrates the insights generative models such as ERGMs could offer by designing simulation experiments for relevant counterfactual questions. In our view, this provides a reminder that network models are not merely statistical hypothesis-testing tools, but flexible and powerful generative devices that can reveal emergent effects of multiple mechanisms on outcomes of interest in complex social systems.

In closing, we note that while statistical network models have seen great advances over the past 20 years, important challenges remain. Among these is the problem of accounting for measurement error (a persistent challenge for the field since the famous call-to-arms of Bernard et al. (1984)). As with the vast majority of work in both social network analysis and demography, this paper considers the data as a fixed input without accounting for measurement error. However, even Census data is imperfectly measured, a concern that becomes greater when considering the $\mathcal{O}(3000^2)$ migration rates that must be estimated to measure the U.S. county-level migration system. Assessing the nature and consequences of measurement error in migration networks remains an open problem, as does the estimation of count-valued ERGMs in the presence of measurement error. These would seem to be important directions for further work.

Likewise, in defining a network, one's choice of nodes and edges imposes a certain level of granularity on one's representation, which in turn impacts what effects it can distinguish (Butts, 2009). Here, we examine the network of migration flows among U.S. counties, which could itself be seen as an aggregation of an ensemble of migrant flow networks for smaller subsets of the U.S. population; although we can hypothesize how these subflows contribute to the aggregate flow network, we are limited in our ability to disaggregate them here. For example, we do not have information about whether and to what extent the larger migration flow from low-White-concentration counties to higher-White-concentration counties is driven by movement of the non-Hispanic White population, versus members of minority populations following on the heels of earlier migration by non-Hispanic Whites (an effect seen in some past research, e.g. Woldoff (2011)). Distinguishing the migration patterns of different population groups within a joint model imposes significant challenges both from a data availability/accuracy and computational standpoint, but could provide further insights if feasible.

Last but not least, we note that there exist other states whose population redistribution patterns are stronger than California, such as New York State and Alaska, despite receiving less public attention. The impacts of the pandemic on internal migration, over both the short term and the long term (e.g., potential enhancement of ex-urban migration), are also critical research topics. We encourage future research to examine these cases to offer a more comprehensive understanding regarding the evolution of the U.S.

migration system and its implications for American society.

References

Almquist, Z. W. and Butts, C. T. (2015). Predicting Regional Self-Identification from Spatial Network Models. *Geographical Analysis*, 47(1):50–72.

Anderson, J. E. (2011). The Gravity Model. *Annual Review of Economics*, 3(1):133–160.

Azose, J. J. and Raftery, A. E. (2015). Bayesian Probabilistic Projection of International Migration. *Demography*, 52(5):1627–1650.

Azose, J. J. and Raftery, A. E. (2018). Estimating large correlation matrices for international migration. *The Annals of Applied Statistics*, 12(2):940–970.

Bahnsen, D. L. (2021). The Great California Exodus. *National Review*, LXXIII(7).

Bakewell, O. (2014). Relaunching migration systems. *Migration Studies*, 2(3):300–318.

Bakewell, O., Engbersen, G., Fonseca, M. L., and Horst, C. (2016). *Beyond Networks: Feedback in International Migration*. Springer.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Beam, A. (2021). California's growth rate at record low as more people leave. *AP NEWS*.

Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J., and Hugo, G. (2002). Cross-national comparison of internal migration: issues and measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(3):435–464.

Bernard, H. R., Killworth, P., Kronenfeld, D., and Sailer, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual review of anthropology*, 13(1):495–517.

Biagi, B., Faggian, A., and McCann, P. (2011). Long and Short Distance Migration in Italy: The Role of Economic, Social and Environmental Characteristics. *Spatial Economic Analysis*, 6(1):111–131. Publisher: Routledge _eprint: https://doi.org/10.1080/17421772.2010.540035.

Boustan, L., Cai, C., and Tseng, T. (2023). JUE Insight: White flight from Asian immigration: Evidence from California Public Schools. *Journal of Urban Economics*, page 103541.

Boyle, P., Keith H., H., Vaughan, R., and Vaughan, R. (2014). *Exploring Contemporary Migration*. Routledge, Abingdon, United Kingdom.

Butts, C. T. (2003). Predictability of large-scale spatially embedded networks. In Breiger, R., Carley, K. M., and Pattison, P., editors, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, pages 313–323. National Academies Press, Washington, D.C.

Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science*, 325:414–416.

Butts, C. T. and Acton, R. M. (2011). Spatial Modeling of Social Networks. In *The SAGE Handbook of GIS and Society*, pages 222–250. SAGE Publications, Inc., 1 Oliver's Yard, 55 City Road London EC1Y 1SP.

Carling, J. (2002). Migration in the age of involuntary immobility: Theoretical reflections and Cape Verdean experiences. *Journal of Ethnic and Migration Studies*, 28(1):5–42.

Carling, J. and Schewel, K. (2018). Revisiting aspiration and ability in international migration. *Journal of Ethnic and Migration Studies*, 44(6):945–963.

Crowder, K., Pais, J., and South, S. J. (2012). Neighborhood Diversity, Metropolitan Constraints, and Household Migration. *American Sociological Review*, 77(3):325–353.

Crowder, K., South, S. J., and Chavez, E. (2006). Wealth, Race, and Inter-Neighborhood Migration. *American Sociological Review*, 71(1):72–94.

de Haas, H. (2010). The Internal Dynamics of Migration Processes: A Theoretical Inquiry. *Journal of Ethnic and Migration Studies*, 36(10):1587–1617.

Dekker, D., Krackhardt, D., and Snijders, T. A. B. (2007). Sensitivity of MRQAP Tests to Collinearity and Autocorrelation Conditions. *Psychometrika*, 72(4):563–581.

DeWaard, J. and Ha, J. T. (2019). Resituating relaunched migration systems as emergent entities manifested in geographic structures. *Migration Studies*, 7(1):39–58.

Dorsey, C. (2021). America's Mass Migration Intensifies As 'Leftugees' Flee Blue States And Counties For Red. *Forbes*.

Dye, T. R. (1990). *American Federalism: Competition Among Governments*. Lexington Books.

Fawcett, J. T. (1989). Networks, Linkages, and Migration Systems. *The International Migration Review*, 23(3):671–680.

Frey, W. H. (1979). Central City White Flight: Racial and Nonracial Causes. *American Sociological Review*, 44(3):425–448.

Henrie, C. J. and Plane, D. A. (2008). Exodus from the California Core: Using Demographic Effectiveness and Migration Impact Measures to Examine Population Redistribution Within the Western United States. *Population Research and Policy Review*, 27(1):43–64.

Hiltzik, M. (2020). California isn't "hemorrhaging" people, but there are reasons for concern. *Los Angeles Times*.

Holland, P. W. and Leinhardt, S. (1981). An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Huang, P. and Butts, C. T. (2022). Rooted America: Immobility and Segregation of the Inter-county Migration Networks. arXiv:2205.02347.

Huang, P. and Butts, C. T. (2024). Parameter estimation procedures for exponential-family random graph models on count-valued networks: A comparative simulation study. *Social Networks*, 76:51–67.

Krackhardt, D. (1988). Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social Networks*, 10(4):359–381.

Krackhardt, D. and Stern, R. N. (1988). Informal Networks and Organizational Crises: An Experimental Simulation. *Social Psychology Quarterly*, 51(2):123–140.

Kritz, M. M., Lim, L. L., Zlotnik, H., and Lim, L. L. L. L. (1992). *International migration systems: a global approach*. Oxford University Press, USA.

Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electronic journal of statistics*, 6:1100–1128.

Kruse, K. M. (2013). White Flight: Atlanta and the Making of Modern Conservatism. In *White Flight*. Princeton University Press.

Leal, D. F. (2021). Network Inequalities and International Migration in the Americas. *American Journal of Sociology*, 126(5):1067–1126.

Lee, E. S. (1966). A theory of migration. *Demography*, 3(1):47–57.

Lichter, D. T. and Brown, D. L. (2011). Rural America in an Urban Society: Changing Spatial and Social Boundaries. *Annual Review of Sociology*, 37(1):565–592.

Liu, X., Andris, C., and Desmarais, B. A. (2019). Migration and political polarization in the U.S.: An analysis of the county-level migration network. *PLOS ONE*, 14(11):e0225405.

Mabogunje, A. L. (1970). Systems Approach to a Theory of Rural-Urban Migration. *Geographical Analysis*, 2(1):1–18.

Partridge, M. D., Rickman, D. S., Olfert, M. R., and Ali, K. (2012). Dwindling U.S. internal migration: Evidence of spatial equilibrium or structural shifts in local labor markets? *Regional Science and Urban Economics*, 42(1-2):375–388.

Patuelli, R. (2016). Spatial Autocorrelation and Spatial Interaction. In *Encyclopedia of GIS*, pages 1–7.

Plane, D. A. (1999). Migration Drift. *The Professional Geographer*, 51(1):1–11.

Plane, D. A., Henrie, C. J., and Perry, M. J. (2005). Migration up and down the urban hierarchy and across the life course. *Proceedings of the National Academy of Sciences*, 102(43):15313–15318.

Poot, J., Alimi, O., Cameron, M. P., and Maré, D. C. (2016). The Gravity Model of Migration: The Successful Comeback of an Ageing Superstar in Regional Science. *SSRN Electronic Journal*.

Preuhs, R. R. (1999). State Policy Components of Interstate Migration in the United States. *Political Research Quarterly*, 52(3):527–549.

Ravenstein, E. G. (1885). The Laws of Migration. *Journal of the Royal Statistical Society*, 48(2):167–235.

Rees, P., Bell, M., Kupiszewski, M., Kupiszewska, D., Ueffing, P., Bernard, A., Charles-Edwards, E., and Stillwell, J. (2017). The Impact of Internal Migration on Population Redistribution: an International Comparison. *Population, Space and Place*, 23(6):e2036.

Shryock, H. S., Siegel, J. S., and Larmon, E. A. (1973). *The Methods and Materials of Demography*. U.S. Bureau of the Census.

Song, S. (2021). Study shows California exodus, with more people leaving the state despite the pandemic. *KTVU FOX 2*.

Tam Cho, W. K., Gimpel, J. G., and Hui, I. S. (2013). Voter Migration and the Geographic Sorting of the American Electorate. *Annals of the Association of American Geographers*, 103(4):856–870.

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

Tolnay, S. E. (2003). The African American "Great Migration" and Beyond. *Annual Review of Sociology*, 29(1):209–232.

Treyz, G. I., Rickman, D. S., Hunt, G. L., and Greenwood, M. J. (1993). The Dynamics of U.S. Internal Migration. *The Review of Economics and Statistics*, 75(2):209–214.

Windzio, M., Teney, C., and Lenkewitz, S. (2019). A network analysis of intra-EU migration flows: how regulatory policies, economic inequalities and the network-topology shape the intra-EU migration space. *Journal of Ethnic and Migration Studies*, 47(5):951–969.

Woldoff, R. A. (2011). *White Flight/Black Flight: The Dynamics of Racial Change in an American Neighborhood*. Cornell University Press.

Zipf, G. K. (1946). The P1 P2/D Hypothesis: On the Intercity Movement of Persons. *American Sociological Review*, 11(6):677–686.

Appendix

Table A1: Valued ERGM for Inter-County Migration Flows, 2011-2015 (Full Model)

	Estimate	Std Err
Key Covariates in Table 2	(included)	
Dependence Structures		
Mutuality	.047***	.002
Waypoint flow	-.013***	.001
Log(past migrant flow)	.300***	<.001
Demographic Covariates		
Origin log(population size)	.353***	.002
Origin log(population size)	.374***	.002
Destination log(population density)	-.081***	.001
Origin log(population density)	-.055***	.001
Destination PSR	.017***	.001
Origin PSR	.017***	.001
Destination log(immigrant inflow)	.057***	.001
Origin log(immigrant inflow)	.043***	.001
Economic Covariates		
Destination P(higher education)	.386***	.012
Origin P(higher education)	.314***	.013
Destination P(renter)	.404***	.012
Origin P(renter)	.421***	.012
Difference P(unemployment)	-1.229***	.041
Origin P(unemployment)	-2.948***	.052
Geographical Covariates		
Log(distance)	-.569***	.001
Same state	.500***	.002
Northeast	(reference group)	
Destination South	.257***	.003
Origin South	.053***	.003
Destination West	.384***	.004
Origin West	.225***	.004
Dstimation Midwest	.202***	.003
Origin Midwest	.096***	.003
Baseline		
sum	-1.421***	.042
nonzero	-13.965***	.028

Note: $^{**}p < 0.01$; $^{***}p < 0.001$ (two-tailed tests).

PNAS BRIEF REPORT | SOCIAL SCIENCES OPEN ACCESS

Geographical patterns of social cohesion drive disparities in early COVID infection hazard

Loring J. Thomas[a], Peng Huang[a,b], Fan Yin[b], Junlan Xu[b], Zack W. Almquist[c,d,e,f,g], John R. Hipp[a,h], and Carter T. Butts[a,b,i,j,1]

Edited by Douglas Massey, Princeton University, Princeton, NJ; received November 30, 2021; accepted January 18, 2022

The uneven spread of COVID-19 has resulted in disparate experiences for marginalized populations in urban centers. Using computational models, we examine the effects of local cohesion on COVID-19 spread in social contact networks for the city of San Francisco, finding that more early COVID-19 infections occur in areas with strong local cohesion. This spatially correlated process tends to affect Black and Hispanic communities more than their non-Hispanic White counterparts. Local social cohesion thus acts as a potential source of hidden risk for COVID-19 infection.

COVID-19 | spatial heterogeneity | diffusion | health disparities | social networks

The spread of COVID-19 has infected millions globally (1) and, in the United States, this has disproportionately affected Black and Latino populations (2). The COVID-19 pandemic has been shown to spread unevenly over social and geographic space (3–5); however, the mechanistic connections between contact network structure and infection hazard are not fully understood. Here, we show that small differences in local social cohesion can result in large disparities in infection rates by race and ethnicity as observed in the United States (6).

While long-term outcomes are important, we specifically aim to understand how the disparities in infection by race and ethnicity arise early in the pandemic. In the initial phase of an emerging pandemic, risks are unclear, nonpharmaceutical interventions (e.g., masking, distancing) are not yet implemented, and behavioral changes are rarely widespread; yet it is precisely at this point that the virus has the greatest opportunity to penetrate the population, with the capacity to provide particular harms to vulnerable communities.

Using a previously published explicit contact network model based on viral dynamics in the early COVID-19 pandemic (3), we examine the network properties that drive differences in initial infection hazard. As Fig. 1 shows, wild-type severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) does not diffuse readily through linear "infection chains" with multiple intermediates; even when multiple, parallel chains connect two individuals, many chains are required to achieve a large infection risk. By contrast, SARS-CoV-2 spreads extremely well through cohesive subgroups, where multiple, redundant ties provide numerous avenues for infection to occur. Being connected to an infective by shared membership in even a fairly small cohesive group results in a dramatic increase in infection risk, due to the factorial increase in the number of potential infection paths with group size. For example, an otherwise isolated susceptible linked to an infective via a clique of only six individuals has a 50% probability of becoming infected; to reach the same infection probability by connection with independent paths of the type shown in Fig. 1 would require maintaining 38 contacts involving 76 intermediaries. This suggests that small differences in social cohesion can lead to large disparities in infection risk for wild-type SARS-CoV-2, much as small differences in partnership concurrency have been shown to drive disparities in HIV risk (7).

To determine whether these network effects would be expected to manifest under realistic conditions, we employ the above model (3) to study early pandemic infection hazards in the city of San Francisco, CA, a major city with a diverse population that suffered significant disparities in pandemic outcomes. We examine the period before March 24, 2020, 1 wk after infection data became available for the four major racial/ethnic groups; by this time, the infection was already spreading throughout the city, and significant racial and ethnic disparities in incidence had emerged. The observed patterns of disparity are typical of what would be expected given the underlying network process, with disparities in infection risks being greatly enhanced by differences in social cohesion. As we further show through simulation, these differences are expected to be geographically correlated, leading to a high-risk "floodplain" that is particularly exposed to infection, and metaphorical "high ground" that is relatively protected.

Author affiliations: [a]Department of Sociology, University of California, Irvine, CA 92697; [b]Department of Statistics, University of California, Irvine, CA 92697; [c]Department of Sociology, University of Washington, Seattle, WA 98195; [d]Department of Statistics, University of Washington, Seattle, WA 98195; [e]Center for Studies in Demography and Ecology, University of Washington, Seattle, WA 98195; [f]Center for Statistics and the Social Sciences, University of Washington, Seattle, WA 98195; [g]eScience Institute, University of Washington, Seattle, WA 98195; [h]Department of Criminology, Law & Society, University of California, Irvine, CA 92697; [i]Department of Computer Science, University of California, Irvine, CA 92697; and [j]Department of Electrical Engineering and Computer Science, University of California, Irvine, CA 92697

Author contributions: Z.W.A., J.R.H., and C.T.B. designed research; L.J.T., P.H., F.Y., J.X., and Z.W.A. performed research; L.J.T. and C.T.B. contributed new reagents/analytic tools; L.J.T., P.H., and J.X. analyzed data; and L.J.T., P.H., Z.W.A., J.R.H., and C.T.B. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: buttsc@uci.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2121675119/-/DCSupplemental.

Published March 14, 2022.

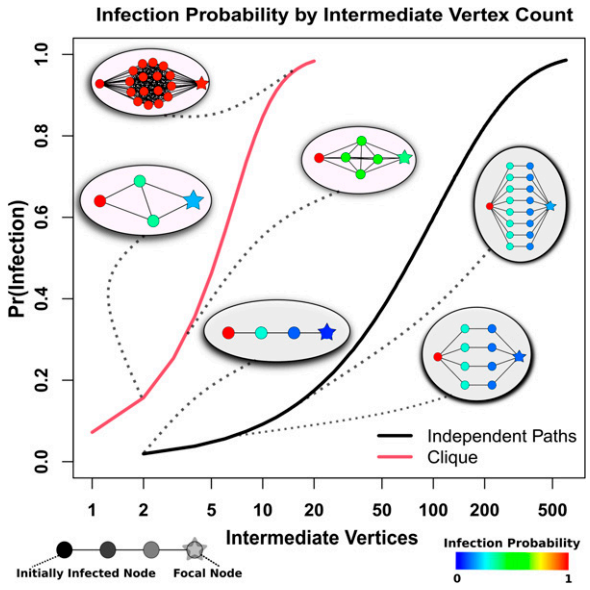


Fig. 1. Probability of diffusion from an infected (*Left*) to uninfected (*Right*) individual bridged by intermediaries arranged in cliques (red curve) versus independent paths (black curve). Comembership in a cohesive subgroup fields infection risks that climb sharply with the number of intermediaries, while much larger numbers of intermediaries are required to obtain the same risk in the case of independent paths.

Results

Infection Outcomes. We simulate 1,225 infection trajectories ("pandemic histories") for the city of San Francisco (*Materials and Methods*) covering the period up to March 24, 2020. Fig. 2*B* shows the resulting distribution of early infection disparities by demographic group (Hispanic [H], non-Hispanic Black [B], non-Hispanic White [W], and non-Hispanic Asian [A]) on March 24, 2020 of the simulation. Because outbreaks can vary greatly in size and timing, early period disparities can and do vary by trajectory. However, we see that Hispanics are hardest hit in the majority of cases, typically followed by Blacks and then Asians. Non-Hispanic Whites are very rarely the hardest hit, and are often (but not always) the group with the lowest early incidence; we note more variability in the identity of the least-hit group, as this outcome is sensitive to chance events (i.e., where early outbreaks occur). The observed pattern based on official data (9) is the third-most common pattern that would be expected, and hence fairly typical of what would be expected given the contact process.

Cohesion Drives Infection Hazard. Fig. 1 shows the risk-enhancing effect of cohesion in isolated subnetworks; this effect generalizes to more-realistic scenarios. A Cox proportional hazards model of infection hazard by core number (a common measure of embeddedness in cohesive groups) confirms a large risk enhancement for local cohesion, with persons in cohesive subgroups facing dramatically higher infection risk over time (Fig. 2*C*); in particular, each unit increment in core number increases infection hazard by ∼30%. Different demographic groups have slightly different levels of cohesion (Fig. 2 *A*, *Inset*). The difference in mean core number between the most cohesive group (Hispanic) and the least (non-Hispanic White) is 1.5, translating to an ∼50% mean risk enhancement; while risk levels vary within all groups, a 9.3% higher share of Hispanic versus non-Hispanic White population has greater than average risk (Fig. 2*A*). Differences in local social cohesion thus provide an important structural basis for disparities in early pandemic outcomes between groups.

Spatial Correlation of Cohesion Produces a Network "Floodplain". Contact network cohesion is spatially correlated, producing areas with higher than average membership in cohesive subgroups, and hence elevated mean risk. Fig. 3*A* shows the mean infection hazard modifier (net of global average) for each US Census block in San Francisco, based on the distribution of cohesion scores (core numbers). Cyan and green areas are epidemiological "high ground" where lower levels of local cohesion reduce mean risk, while red and orange areas are epidemiological "floodplains" where high cohesion leads to enhanced local risk. These cohesion-driven patterns are well correlated with the overall rate of infections, as illustrated by the mean inverse infection time across the city (Fig. 3*B*). Spatial segregation in housing places some groups in harm's way, increasing disparities in incidence during the initial outbreak.

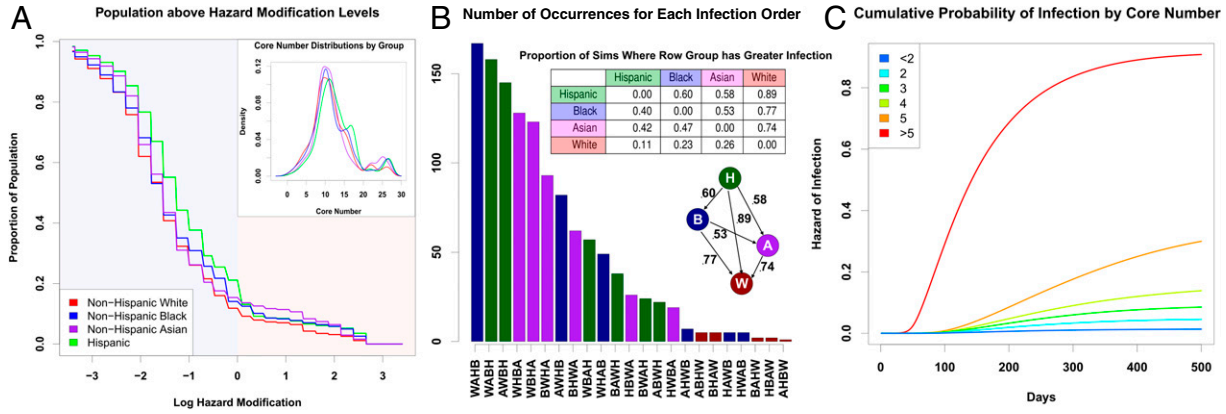


Fig. 2. (*A*) Proportion of each population that lives "below" a given point on the floodplain (higher risk), denoted by its log hazard modification. The non-Hispanic White population is consistently present on the higher parts of the floodplain, with the non-Hispanic Asian population also being present in the middle of the floodplain. The lower parts of the floodplain are heavily occupied by non-Hispanic Black and Hispanic populations. (*Inset*) Distribution of core numbers for each ethnoracial group in the San Francisco model; small differences in core numbers are sufficient to drive large differences in risk. (*B*) Distribution of qualitative outcomes in simulation on March 24, where *x* axis labels correspond to group labels in order of infection rates, from lowest (bottom) to highest (top) prevalence. Bars are colored corresponding to the group with highest prevalence. The third bar (order AWBH) corresponds to the observed pattern from San Francisco. (*Top Inset*) The proportion of times each row group has a greater infection rate than the column group across all simulations. The Hispanic population consistently has the highest infection rates, followed, on average, by the Black population, the Asian population, and the non-Hispanic White population. (*Bottom Inset*) A graph describing the proportion of simulations one group (tail) has a greater infection rate than another (head). (*C*) Cumulative probability of infection by core number from simulated networks. Higher core numbers indicate greater levels of local cohesion, which substantially increases one's hazard of infection. The bicomponent, where core number is equal to two, does not seem to drive infection patterns, as some prior literature suggests (8).

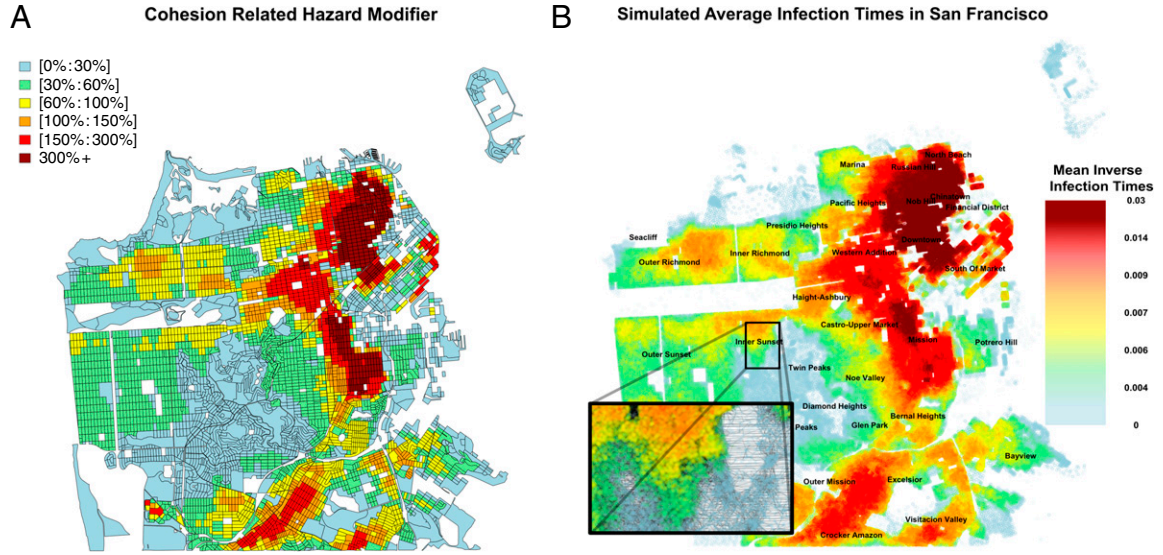


Fig. 3. (*A*) Average deviation from the mean hazard attributable to core number, across San Francisco. Risk enhancement is spatially correlated, with significant risk downtown and much lower risk near the central part of the city. These hazards form a "floodplain," where some areas are more dangerous than others. (*B*) Simulated infection times across San Francisco, averaged across 35 simulations. The patterns of infections match the expected hazard modifications in *A*. *Inset* shows the structure of the social network in the Inner Sunset neighborhood.

Discussion

The mere presence of connecting paths is not sufficient for rapid diffusion of a disease like wild-type SARS-CoV-2: Infection of contacts is rare enough to require considerable redundancy for transmission to occur. Cohesion greatly increases the number of potential infection pathways, rendering an otherwise relatively "opaque" network "transparent" to disease transmission. The uneven distribution of cohesive subgroups in large networks and their much greater permeability help to explain the "bursty" nature of SARS-CoV-2 diffusion, with slow diffusion through less cohesive parts of the network punctuated by rapid outbreaks in cohesive groups (3, 10). Ironically, social cohesion has long been viewed as a community asset, particularly with respect to community resilience following disasters or other sources of social disruption (11–13); in the context of an infection like SARS-CoV-2, this same cohesion can act as an epidemiological risk factor. Local cohesion varies by location, with some parts of the San Francisco network having higher local cohesion than others. Combined with high levels of residential segregation, these differences can, in turn, produce disparities in infection hazard by race and ethnicity. In San Francisco, we find that Black and Hispanic populations are expected to have the highest infection rates in the early pandemic, followed by the Asian population and the White non-Hispanic population. Our models suggest that the exact evolution of infection rates is somewhat contingent on chance events, and multiple scenarios are possible based on which subgroups are hit first; however, some scenarios are muchmore likely than others, with the observed pattern of infection in the early pandemic being one of those predicted to be most likely to occur. Greater attention to cohesion as a risk factor—particularly given its spatial correlation—may help to prioritize warning messages or interventions for high-risk groups when outbreaks of a potentially serious disease are first detected.

Materials and Methods

Population data for the COVID-19 simulation are from 2010 block-level US Census data for San Francisco. The number of observed infection cases of each racial group comes from San Francisco Department of Public Health (9). Contact network simulations and COVID-19 transmission employ the published model of ref. 3, with additional corrections for recovery and mortality hazards by age and sex as well as the date of the existence of patient 0, as described in *SI Appendix*. Model and parameterization details are contained in *SI Appendix*, along with the simulation details. Assessment of the cohesion/infection hazard relationship was performed via Cox proportional hazards models; parameterization details are provided in *SI Appendix*. Cross-tabulation of expected risk enhancement by areal unit and group produced the results of Figs. 2*A* and 3.

Data Availability. R objects containing spatial Bernoulli networks and code for analysis of simulated network data have been deposited in Harvard Dataverse (https://doi.org/10.7910/DVN/NT4KDH) (14).

ACKNOWLEDGMENTS. This work was supported by NSF Awards IIS-1939237 and SES-1826589 to C.T.B., NIH Award P2C HD042828 to the Center for Studies in Demography and Ecology for Z.W.A., and a University of California, Irvine Council on Research, Computing and Libraries grant.

1. Centers for Disease Control and Prevention, CDC COVID data tracker. https://covid.cdc.gov/covid-data-tracker/#datatracker-home. Accessed 1 February 2021.
2. T. Andrasfay, N. Goldman, Reductions in 2020 US life expectancy due to COVID-19 and the disproportionate impact on the Black and Latino populations. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2014746118 (2021).
3. L. J. Thomas *et al.*, Spatial heterogeneity can lead to substantial local variations in COVID-19 timing and severity. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 24180–24187 (2020).
4. B. Hong, B. J. Bonczak, A. Gupta, L. E. Thorpe, C. E. Kontokosta, Exposure density and neighborhood disparities in COVID-19 infection risk. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2021258118 (2021).
5. X. Hou *et al.*, Intracounty modeling of COVID-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2020524118 (2021).
6. Centers for Disease Control and Prevention, COVID-19 hospitalization and death by race/ethnicity. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html. Accessed 1 February 2021.
7. M. Morris, H. Epstein, M. Wawer, Timing is everything: International variations in historical sexual partnership concurrency and HIV prevalence. *PLoS One* **5**, e14092 (2010).
8. J. Moody, J. Adams, M. Morris, Epidemic potential by sexual activity distributions. *Netw. Sci. (Camb. Univ. Press)* **5**, 461–475 (2017).
9. San Francisco Department of Public Health, COVID-19 cases summarized by race and ethnicity. https://data.sfgov.org/COVID-19/COVID-19-Cases-Summarized-by-Race-and-Ethnicity/vqqm-nsqg. Accessed 21 April 2021.
10. F. Wong, J. J. Collins, Evidence that coronavirus superspreading is fat-tailed. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29416–29418 (2020).
11. C. Fan, Y. Jiang, A. Mostafavi, Emergent social cohesion for coping with community disruptions in disasters. *J. R. Soc. Interface* **17**, 20190778 (2020).
12. I. Townshend, O. Awosoga, J. Kulig, H. Fan, Social cohesion and resilience across communities that have experienced a disaster. *Nat. Hazards* **76**, 913–938 (2015).
13. J. E. Cinner *et al.*, Sixteen years of social and ecological dynamics reveal challenges and opportunities for adaptive management in sustaining the commons. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 26474–26483 (2019).
14. T. J. Loring et al., Geographical Patterns of Social Cohesion Drive Disparities in Early COVID Infection Hazard. Harvard Dataverse. https://doi.org/10.7910/DVN/NT4KDH. Deposited 4 March 2022.

PNAS
www.pnas.org

Supplementary Information for

Geographical Patterns of Social Cohesion Drive Disparities in Early COVID Infection Hazard

Loring J. Thomas, Peng Huang, Fan Yin, Junlan Xu, Zack W. Almquist, John R. Hipp, Carter T. Butts

Carter T. Butts.
E-mail: buttsc@uci.edu

This PDF file includes:

Supporting Information Text

Introduction

In this appendix, we include additional information about the parameterization of the diffusion model, as well as the Cox Proportional-Hazards model. This section also provides more detail on the data that is used to generate the networks and estimate parameters.

Network and Demographic Data

We employ data from the 2010 U.S. Census to generate the population level social networks that underlie the analysis in this manuscript. Specifically, we use the smallest level of geography publicly available from the U.S. Census, known as the U.S. Census block level (approximately a city block in an urban setting). Each block contain basic demographic information, including household size.

To generate the network, we employ spatial network models that rely on a kernel function (the Spatial Interaction Functions, SIFs) to describe the presence of a social tie based on the distance between nodes; each node represents a single individual, and all simulations explicitly track the infection history of each individual in the population (as well as their infection paths). We employ the same network generation process used by Thomas et al. (1), which leverages the strategy of (2, 3) of placing households within Census blocks using a low-discrepancy (Halton) sequence, followed by jittered placement of individual locations about the household center. To parameterize the model used in this manuscript, we need to first define the spatial network models (or spatial Bernoulli models) which depend on the SIF. The SIF describes the probability of a tie being present between any two entities, given the distance between those entities. We use the same SIFs as in Thomas et al. (1) which employ a power law model of the form, $\mathcal{F}(\mathcal{D}_{ij}, \theta) = \frac{p_b}{(1+\alpha\mathcal{D}_{ij})^\gamma}$, where p_b describes the baseline probability of a tie existing, α is a scaling parameter describing the effect of a unit of distance, $\mathcal{D}_{ij}$ is the distance a dyad spans, and γ is a parameter describing the form of the tie probability decay. The simulation process employed uses two SIFs, based on prior literature to generate networks. The parameters for these SIFs can be found in (1).

Departing from prior work, we also leverage demographic information on U.S. Census blocks. These demographic covariates are race, ethnicity, age, and sex. These demographic covariates were assigned to nodes such that the three way distribution of race/sex/age and the two way distribution of race/ethnicity match the observed data at the block level. This allows a more fine-grained parameterization for simulation of the diffusion of COVID across social contact networks, based on demographic characteristics of each node (as detailed in the next section). We note that our procedure also leverages household size and thus represents the increased likelihood of being in a clique for individuals in such settings. This factor is one of the core factors that leads to COVID risk, as household spread of the disease is a primary avenue of spread.

We apply this technique to map social contact networks of San Francisco for three core reasons. (i) San Francisco is a city/county administrative unit – this is important because most data reported for the COVID-19 pandemic is at the county level in the U.S. and this allows us to analyze a complete city. (ii) San Francisco is a peninsula that is separated on three sides by water, reducing boundary effects from contacts outside the border of the city. (iii) The city/county of San Francisco published longitudinal data on infections by ethno-racial groups of the early pandemic (4). The combination of good data management and reporting makes San Francisco unique, and when taken together with its status as a natural reporting unit (i.e. also being a county) it becomes an important unique case for studies such as the one conducted in this manuscript. We observe that future decisions by other municipalities to publish longitudinal data broken down by demographics would facilitate further studies of this kind.

In general the epidemiological literature has shown that population density increases the rate of disease spread (5, 6), but it does not provide a mechanistic interpretation for this phenomenon. However, previous research on spatial network models has highlighted the way in which density can drive tie creation and resulting cohesive subgroup formation (3). Our model provides a specific mechanism for how population density and household size distributions may result in increased disease spread: population distribution influences the creation of locally cohesive regions within the contact network, and these regions are *exceptionally permeable* to SARS-CoV-2. It is important to observe that this is not equivalent to number of contacts *per se* - as shown in Fig. 1, susceptibles with large numbers of contacts may still have relatively low infection hazard, when not embedded in a highly cohesive group.

Parameterization of Diffusion Model

To simulate the spread of COVID across a social contact network, we use a continuous time diffusion model defined by (1). This diffusion model describes the way that individuals in the social network experience the disease and spread it to others. This diffusion model begins with the network structure and a vector of disease states for each node (individual). Disease states can be Susceptible (an individual who does not have the disease, but can get infected with it), Infected (the individual has been infected with the disease, but is not infectious), Infectious (the individual can spread the disease to others), Dead, or Recovered. At the beginning of the simulation, all nodes begin in the Susceptible state, with the exception of the seed infections. These nodes begin the simulation being infected with the disease. 25 individuals, randomly selected from the population, are the seed infections in each of the simulations.

Simulations are run until a steady state has been achieved, in which there are no more infected or infectious people, with everyone being in the Susceptible, Recovered, or Dead states. At this point, the diffusion model provides a detailed history

for each node, describing the individual's final state in the simulation, as well as the times at which the node entered any given state. The disease spreads across the structure of the network, with connected nodes being able to transmit the disease across their social ties. Infection occurs as a Poisson event with a fixed rate, described by (1). Only infectious nodes can infect susceptible social contacts; once an individual recovers or dies, they are no longer able to infect or be infected with COVID. When a Susceptible node is infected by an infectious alter, a Bernoulli trial is performed, determining whether a node becomes terminally or non-terminally infected. The rate of success (terminal infection) of the Bernoulli trial is given by P_d, a matrix sorted by age in the row and sex in the column (top to bottom row: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80+; left to right column: female and male); for an individual with age category i and sex category j, the indicator for terminal infection thus arises as $T_{ij} \sim \text{Bern}(P_{d_{ij}})$. P_d, which is in essence a transformation of the Infection Fatality Ratio (IFR) broken down by age and sex, is calculated based on two pieces of information: the IFR for each age group (7), and the sex ratio of death probability within each age group (8), assuming the probability of male and female getting infected is equal within each age group. P_d describes the set of Bernoulli parameters determining the likelihood of a fatal infection:

$$P_d = \begin{pmatrix} 0.000022 & 0.000018 \\ 0.000049 & 0.000049 \\ 0.000216 & 0.000384 \\ 0.000604 & 0.000996 \\ 0.001045 & 0.001955 \\ 0.003625 & 0.008375 \\ 0.012360 & 0.036410 \\ 0.030357 & 0.071643 \\ 0.070189 & 0.115811 \end{pmatrix}$$

The timing of transitions between different states is governed by a series of Gamma distributions. The waiting time from being infected to being infectious is governed by a Gamma distribution with shape 5.807 and scale 0.948, as estimated by (9). For transition towards recovery or death, while prior work used homogeneous distributions, we break them down by demographics to more accurately account for variation across different populations. We estimate their parameters by matching the mean and standard deviation of waiting time for each group, using epidemiological data reported in (10–12). These method of moments estimators coincide with maximum likelihood estimators for the associated parameters, given that the Gamma distribution is a member of the exponential family. Specifically, the waiting time to death for a terminally infected individual in age category i is distributed as ${t^d}_i \sim \text{Gamma}(G_{di1}, G_{di2})$, where G_d is a parameter matrix whose columns contain shape and rate parameters, respectively, and rows indicate age category (top to bottom: 0-49, 50-64, and 65+). (Note that we do not vary the waiting time distribution by sex, as we are not aware of applicable time-to-mortality data from the early pandemic that supports age/sex decomposition.) Here, G_d is given as follows:

$$G_d = \begin{pmatrix} 3.744 & 0.251 \\ 3.568 & 0.233 \\ 2.881 & 0.223 \end{pmatrix}$$

The waiting time to recovery is broken down by both age and sex. For a male in age category i with a non-terminal infection, the waiting time to recovery is distributed as $t^r_{im} \sim \text{Gamma}(G^{rm}_{i1}, G^{rm}_{i2})$, where G^{rm} is a parameter matrix whose rows are ordered by age category (top to bottom: 0-19, 20-29, 30-39, 40-49, 50-59, 60+) and whose columns respectively contain shape and rate parameters. Here, G^{rm} is given as follows:

$$G^{rm} = \begin{pmatrix} 5.339 & 0.392 \\ 5.782 & 0.414 \\ 5.808 & 0.402 \\ 6.686 & 0.452 \\ 6.301 & 0.425 \\ 6.242 & 0.424 \end{pmatrix}$$

For a non-terminally infected female in age category i, the waiting time to recovery is similarly distributed as $t^r_{if} \sim \text{Gamma}(G^{rf}_{i1}, G^{rf}_{i2})$, where G^{rf} is a second parameter matrix whose rows are also ordered by age category (top to bottom: 0-19, 20-29, 30-39, 40-49, 50-59, 60+) and whose columns respectively contain shape and rate parameters. G^{rf} is as follows:

$$G^{rf} = \begin{pmatrix} 5.395 & 0.408 \\ 5.623 & 0.402 \\ 5.326 & 0.376 \\ 6.258 & 0.424 \\ 5.776 & 0.407 \\ 4.719 & 0.337 \end{pmatrix}$$

Since the diffusion process precedes the reporting of the first confirmed positive case, we performed a grid search to determine the length of the time lag between the appearance of "patient zero" in the city and the report of the first positive confirmed

case (March 3, 2021 (13)). Our search was performed over an interval from a minimum of 1 and a maximum of 100 days. For each possible number, we regressed the number of infection case for each racial group in their observed time period using data from (4), on its counterparts in the simulation. The loss function is the summation of the mean squared errors (MSE) for all the linear regressions. We find that a 35 day lag minimizes the MSE, and this value is used here.

Simulation Details

Given the network and diffusion models described above, we run a series of simulations in which the population of San Francisco is seeded with randomly placed infectives 35 days prior to the first confirmed case report in San Francisco on March 3, 2021, and the infection process is followed until the end of our observation period (March 24, 2020, one week after demographic data becomes available for all four major racial/ethnic groups within the city). 35 individual-level contact networks were generated for San Francisco, using different simulated node locations for each realization. For each of these 35 simulated networks, we run 35 diffusion replicates, reseeding the seed infections for each simulation. This produces 1225 simulation replicates.These networks were produced with the R programming language, using the `sna` library (14, 15). For results reported about a single network realization in the main text, we average the infection time (or inverse infection time) for each diffusion replicate simulated in that network. The network being averaged across was selected as the network that most closely matches the average infection and susceptibility splits across all networks on March 24, 2020. For other metrics (such as the reported Cox model results), we average across the entire sample of networks. All figures from the main text utilize simulated data calibrated to observed data on infections and deaths.

The number of replications (independently simulated networks and diffusion simulations within network) was chosen based on a preliminary power analysis based on pilot simulations. Due to the diffusion simulation being bound to the structure of the social network, multiple network replicates were used to highlight trends in infection patterns across space. Likewise, given that the pandemic trajectories are dependent on the seed locations in the network, we randomized the seeds in each pandemic replicate to ensured that simulated trends were not due to idiosyncrasies in seed placement in the network structure. (The equality between the replication count and the inferred optimal lag time for the first infection is coincidental.)

Cox Proportional-Hazards Models

To assess the effects of local cohesion on infection hazards, we use Cox Proportional-Hazards models. Cox models control for (possibly time-varying) background hazards, allowing us to identify the impact of cohesion on infection hazard net of the overall progress of the outbreak. Because each simulated outbreak follows a distinct trajectory, we fit a single model to each simulated trajectory (with the baseline hazard, plus a single effect for core number). This model predicts the hazard of an uninfected individual getting infected with COVID-19, using the core number of a given node (16) as a cohesion measure. The *core number* of a node - specifically, the highest k such that the node belongs to the kth degree core of the contact network - is a measure of local cohesion, with higher numbers indicating that the focal node is embedded in a more cohesive subgroup. In particular, nodes with core numbers of 0 are isolates, those of core 1 belong to trees or pendant trees, and those of core number 2 or higher belong to bicomponents (with higher numbers indicating higher levels of of cohesion). The core number is measured in units of ties, with a core number of k indicating that ego has at least k ties to alters who themselves have core numbers of at least k (and hence who have at least k ties to others with at least k ties to others in the core, recursively). We note that core number is not equivalent to degree: one can have arbitrarily high degree and still have a core number as low as 1. The Cox model coefficient for core number thus indicates the extent to which nodes embedded in locally cohesive regions within the contact network are infected more or less rapidly (on average) than other nodes, controlling for the time-varying baseline infection hazard.

The form of the Cox used here is $h(t) = h_b(t)\exp(\beta X)$. Here, $h(t)$ represents the infection hazard, with $h_b(t)$ being the baseline hazard, X the core number, and β a coefficient expressing the increase in the log infection hazard per unit increase in core number. Here, we observed a mean β of 0.2615 over all simulations, implying an average risk enhancement of approximately 30% in infection hazard per unit increase in core number (as reflected in Fig.2C). As described in the main text, cohesion is a strong and consistent risk factor for early COVID infection, with nodes in high-order cores having a much higher infection risk than those in low-order cores.

Code and Data Availability

We have provided the code and data used for this project, including all parameters for the demographic models. This archive can be found at https://doi.org/10.7910/DVN/NT4KDH.

References

1. Thomas LJ, et al. (2020) Spatial Heterogeneity can Lead to Substantial Local Variations in COVID-19 Timing and Severity. *Proceedings of the National Academy of Sciences* 117(39):24180–24187.
2. Almquist ZW, Butts CT (2012) Point Process Models for Household Distributions Within Small Areal Units. *Demographic Research* 26:593–632.
3. Butts CT, Acton RM, Hipp JR, Nagle NN (2012) Geographical variability and network structure. *Social Networks* 34:82–100.

4. San Francisco Department of Public Health (2021) COVID-19 Cases Summarized by Race and Ethnicity (https://data.sfgov.org/COVID-19/COVID-19-Cases-Summarized-by-Race-and-Ethnicity/vqqm-nsqg). Accessed: 4/21/2021.
5. Kadi N, Khelfaoui M (2020) Population density, a factor in the spread of COVID-19 in Algeria: Statistical study. *Bulletin of the National Research Centre* 44(1):1–7.
6. Rashed EA, Kodera S, Gomez-Tames J, Hirata A (2020) Influence of absolute humidity, temperature and population density on COVID-19 spread and decay durations: Multi-prefecture study in Japan. *International Journal of Environmental Research and Public Health* 17(15):5354.
7. Ferguson, Neil and Laydon, Daniel and Nedjati Gilani, Gemma and Imai, Natsuko and Ainslie, Kylie and Baguelin, Marc and Bhatia, Sangeeta and Boonyasiri, Adhiratha and Cucunuba Perez, ZULMA and Cuomo-Dannenburg, Gina and others (2020) Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand (https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf). Accessed: 11/18/2021.
8. Bhopal SS, Bhopal R (2020) Sex Differential in COVID-19 Mortality Varies Markedly by Age. *Lancet (London, England)*.
9. Lauer, Stephen A and Grantz, Kyra H and Bi, Qifang and Jones, Forrest K and Zheng, Qulu and Meredith, Hannah R and Azman, Andrew S and Reich, Nicholas G and Lessler, Justin (2020) The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine* 172(9):577–582.
10. Voinsky I, Baristaite G, Gurwitz D (2020) Effects of Age and Sex on Recovery from COVID-19: Analysis of 5769 Israeli Patients. *Journal of Infection* 81(2):e102–e103.
11. Khalili M, et al. (2020) Epidemiological Characteristics of COVID-19: a Systematic Review and Meta-analysis. *Epidemiology & Infection* 148.
12. CDC (2020) CDC COVID-19 Pandemic Planning Scenarios (https://cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html). Accessed: 9/7/2020.
13. San Francisco Department of Public Health (2021) COVID-19 Cases Over Time (https://data.sfgov.org/COVID-19/COVID-19-Cases-Over-Time/gyr2-k29z). Accessed: 10/07/2021.
14. Butts CT (2008) Social Network Analysis with sna. *Journal of Statistical Software* 24(6):1–51.
15. R Core Team (2013) R: A Language and Environment for Statistical Computing.
16. Seidman SB (1983) Network structure and minimum degree. *Social Networks* 5:269–287.

Spatial heterogeneity can lead to substantial local variations in COVID-19 timing and severity

Loring J. Thomas[a], Peng Huang[a], Fan Yin[b], Xiaoshuang Iris Luo[c], Zack W. Almquist[d], John R. Hipp[c], and Carter T. Butts[a,b,e,f,1]

[a]Department of Sociology, University of California, Irvine, CA, 92697; [b]Department of Statistics, University of California, Irvine, CA, 92697; [c]Department of Criminology, Law, and Society, University of California, Irvine, CA, 92697; [d]Department of Sociology, Center for Studies in Demography and Ecology, Center for Statistics and Social Sciences, eScience, University of Washington, Seattle, WA, 98195; [e]Department of Computer Science, University of California, Irvine, CA, 92697; and [f]Department of Electrical Engineering and Computer Science, University of California, Irvine, CA, 92697

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved August 18, 2020 (received for review June 6, 2020)

Standard epidemiological models for COVID-19 employ variants of compartment (SIR or susceptible–infectious–recovered) models at local scales, implicitly assuming spatially uniform local mixing. Here, we examine the effect of employing more geographically detailed diffusion models based on known spatial features of interpersonal networks, most particularly the presence of a long-tailed but monotone decline in the probability of interaction with distance, on disease diffusion. Based on simulations of unrestricted COVID-19 diffusion in 19 US cities, we conclude that heterogeneity in population distribution can have large impacts on local pandemic timing and severity, even when aggregate behavior at larger scales mirrors a classic SIR-like pattern. Impacts observed include severe local outbreaks with long lag time relative to the aggregate infection curve, and the presence of numerous areas whose disease trajectories correlate poorly with those of neighboring areas. A simple catchment model for hospital demand illustrates potential implications for health care utilization, with substantial disparities in the timing and extremity of impacts even without distancing interventions. Likewise, analysis of social exposure to others who are morbid or deceased shows considerable variation in how the epidemic can appear to individuals on the ground, potentially affecting risk assessment and compliance with mitigation measures. These results demonstrate the potential for spatial network structure to generate highly nonuniform diffusion behavior even at the scale of cities, and suggest the importance of incorporating such structure when designing models to inform health care planning, predict community outcomes, or identify potential disparities.

COVID-19 | spatial heterogeneity | diffusion | health disparities | social networks

Since its emergence at the end of 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread rapidly to all portions of the globe, infecting over 20 million people as of mid-August 2020 (1). The disease caused by this virus, denoted COVID-19, generally manifests as a respiratory illness that is spread primarily via airborne droplets. While most cases of COVID-19 are nonfatal, a significant fraction of those infected require extensive supportive care, and the mortality rate is substantially higher than more common infectious diseases such as seasonal influenza (2). Even for survivors, infection can lead to long-term damage to the lungs and other organs, leading to long convalescence times and enhanced risk of secondary complications (3, 4). By early March of 2020, COVID-19 outbreaks had appeared on almost every continent, including significant clusters within many cities (5). In the absence of an effective vaccine, public health measures to counteract the pandemic in developed nations have focused on social distancing measures that seek to slow diffusion sufficiently to avoid catastrophic failure of the health care delivery system. Both the planning and public acceptance of such measures have been highly dependent upon the use of epidemiological models to probe the potential impact of distancing interventions, and to anticipate when such measures may be loosened with an acceptable level of public risk. As such, the assumptions and behavior of COVID-19 diffusion models are of significant concern.

Currently, dominant approaches to COVID-19 modeling (6–8) are based on compartment models (often called SIR models, after the conventional division of the population into susceptible, infected, and recovered groups in the most basic implementations) that implicitly treat individuals within a population as geographically well mixed. While some such models include differential contact by demographic groups (e.g., age), and may treat states, counties, or, occasionally, cities as distinct units (e.g., work by ref. 9), those models presently in wide use do not incorporate spatial heterogeneity at local scales (e.g., within cities). Past work, however, has shown evidence of substantial heterogeneity in social relationships at regional, urban, and suburban scales (10–12), with these variations in social network structure impacting outcomes as diverse as regional identification (13), disease spread (14), crime rates (15), neighborhood identification, and development (12, 16). If individuals are not socially "well-mixed" at local scales, then it is plausible that diffusion of SARS-CoV-2 via interpersonal contacts will likewise depart from the uniform mixing characteristic of the SIR models. Indeed, at least one computational study (17) using a fairly "generic" (non-COVID) diffusion process on realistic urban networks has shown considerable nonuniformity in diffusion times, suggesting that such effects could hypothetically be present. Variations across local regions on the pandemic timing, severity,

Significance

We examine the effects of an uneven population distribution on the spread of the COVID-19 disease spread, using a diffusion model based on interpersonal contact networks. Taking into account spatial heterogeneity, the spread of COVID-19 is much "burstier" than in standard epidemiological models, with substantial local disparities in timing and severity and long lags between local outbreaks. We show that spatial heterogeneity may produce dramatic differences in social exposures to those with the illness, and may stress health care delivery systems in ways that are not well captured by standard SIR-like models.

Author contributions: J.R.H. and C.T.B. designed research; L.J.T., P.H., F.Y., and Z.W.A. performed research; L.J.T., P.H., Z.W.A., and C.T.B. contributed new reagents/analytic tools; L.J.T., P.H., F.Y., X.I.L., and Z.W.A. analyzed data; and L.J.T., P.H., F.Y., X.I.L., Z.W.A., J.R.H., and C.T.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: buttsc@uci.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2011656117/-/DCSupplemental.

First published September 10, 2020.

and the hospital load could have huge impacts on the social outcomes of different population groups (e.g., racial/ethnic groups) in the pandemic, given the heterogeneity of their spatial distribution in urban and suburban areas (18, 19). However, it could also be hypothesized that such effects would be small perturbations to the broader infection curve captured by conventional compartment models, with little practical importance. The question of whether these effects are likely to be present for COVID-19, and, if so, their strength and size, has, to date, remained open.

In this paper, we examine the potential impact of local spatial heterogeneity on COVID-19, modeling the diffusion of SARS-CoV-2 in populations whose contacts are based on spatially plausible network structures. We focus here on the urban context, examining 19 different cities in the United States. We simulate the population of each city in detail (i.e., at the individual level), simulating hypothetical outbreaks on the contact network in each city in the absence of measures such as social distancing. Despite allowing the population to be well mixed in all other respects (i.e., not imposing mixing constraints based on demographic or other characteristics), we find that spatial heterogeneity alone is sufficient to induce substantial departures from spatially homogeneous SIR behavior. Among the phenomena observed are "long lag" outbreaks that appear in previously unharmed communities after the aggregate infection wave has largely subsided, frequently low correlations between infection timing in spatially adjacent communities, and distinct subpatterns of outbreaks found in some urban areas that are uncorrelated with the broader infection pattern. Gaps between infection peaks at the intraurban level can be large, for example, on the order of weeks or months in extreme cases, even for communities that are within kilometers of each other. Such heterogeneity is potentially consequential for the management of health care delivery services: As we show, using a simple "catchment" model of hospital demand, local variations in infection timing can easily overload hospitals in some areas, generating "hospital deserts" (20), while leaving others relatively empty (absent active reallocation of patients). Likewise, we show that individuals' social exposures to others who are morbid or deceased vary greatly over the course of the pandemic, potentially leading to differences in risk assessment and bereavement burden for persons residing in different locations. Differences in outbreak timing and severity may exacerbate health disparities (since, e.g., surge capacity varies by community) and may even affect perception of and support for prophylactic behaviors among the population at large, with those in so-far untouched communities falsely assuming that the pandemic threat is either past or was exaggerated to begin with, or attributing natural variation in disease timing to the impact of health interventions.

We note at the outset that the models used here are intended to probe the hypothetical impact of spatial heterogeneity on COVID-19 diffusion within particular scenarios, rather than to produce high-accuracy predictions or forecasts. For the latter applications, it is desirable to incorporate many additional features that are here simplified to facilitate insight into the phenomenon of central interest. In particular, we do not incorporate either demographic effects or social distancing (21, 22), allowing us to consider a setting that is as well mixed as possible (and hence as close as possible to an idealized SIR model), with the exception of spatial heterogeneity. As we show, even this basic scenario is sufficient to produce large deviations from the SIR model. Despite the simplicity of our models, we do note that the approach employed here could be integrated with other factors and calibrated to produce models intended for forecasting or similar applications.

Materials and Methods

Spatial Network Data. Networks are generated using population distributions from the most recent US Census in 2010. Network construction followed the same methodology as Butts et al. (23). Hospital information was obtained from the Homeland Infrastructure Foundation-Level Data (HIFLD) database (24). HIFLD is an initiative that collects geospatial information on critical infrastructure across multiple levels of government. We employ the national-level hospital facility database, which contains locations of hospitals for the 50 US states; Washington, DC; and US territories of Puerto Rico, Guam, American Samoa, Northern Mariana Islands, Palau, and Virgin Islands; underlying data are collated from various state departments or federal sources (e.g., Oak Ridge National Laboratory). We employ all hospitals within our 19 target cities, excluding facilities closed since 2019. Latitude/longitude coordinates and capacity information were employed to create a spatial database that includes information on the number of beds in each hospital. The capacity information includes the number of beds that each hospital has available, and can be used to assess strain that a surge in hospitalizations could create.

The dates of the first confirmed case and all of the death cases for King County, where Seattle is located, were obtained from *The New York Times*, based on reports from state and local health agencies (25). The death rate was calculated based on population size of each county from the 2018 American Community Survey, and employed to calibrate the infection rate (the only free parameter in the models used here); details are provided in *SI Appendix*.

We ran 10 replicates of the COVID-19 diffusion process in each of our 19 cities, seeding with 25 randomly selected infections in each replicate and following the course of the diffusion until no infectious individuals remained. Simulations were performed using a combination of custom scripts for the R statistical computing system (26) and the statnet library (27–29). Analyses were performed using R.

Methods. COVID-19 is typically transmitted via direct contact with infected individuals, with the greatest risk occurring when an uninfected person is within approximately six feet of an infected person for an extended period. Such interactions can be modeled as events within a social network, where individuals are tied to those with whom they have a high hazard of intensive interaction. In prior work, this approach has been successfully employed for modeling infectious diseases ranging from HIV (30) and influenza (31) to Zika (32). To model networks of potential contacts at scale, we employ spatial network models (33), which are both computationally tractable and able to capture the effects of geography and population heterogeneity on network structure (23). Such models have been successfully used to capture social phenomena ranging from neighborhood-level variation in crime rates (15) and regional identification (13) to the flow of information among homeless persons (34).

The spatial network models used here allow for complex social dependence through a kernel function, referred to as the *social interaction function* (SIF). The SIF formally defines the relationship between two individuals based on spatial proximity. For example it has been shown that many social interaction patterns obey the Zipf law (35), where individuals are more likely to interact with others close by rather than far away [a pattern that holds even for online interactions (10)]. Here, we use this approach to model a network that represents the combination of frequent interactions due to ongoing social ties and contacts resulting from frequent incidental encounters (e.g., interactions with neighbors and community members).

We follow the protocol of refs. 15 and 23 to simulate social network data that combine the actual distribution of residents in a city with a prespecified SIF. We employ the model of ref. 15 with decennial Census data to produce large-scale social networks for 19 cities and counties in the United States—providing a representation of major urban areas in the United States (*SI Appendix*). Given these simulated networks, we then implement an individual-level SIR-like framework to examine COVID-19 diffusion. At each moment in time, each individual can be in a susceptible, infected but not infectious, infectious, deceased, or recovered state. The disease diffuses through the contact network, with currently infectious individuals infecting susceptible neighbors as a continuous time Poisson process with a rate estimated from mortality data (*SI Appendix*); recovered or deceased individuals are not considered infectious for modeling purposes. Upon infection, an individual's transitions between subsequent states (and into mortality or recovery) are governed by waiting time distributions based on epidemiological data as described in *SI Appendix*. To begin each simulated trajectory, we randomly infect 25 individuals, with all others being considered susceptible. Simulation proceeds until no infectious individuals remain.

From the simulated trajectory data, we produce several metrics to assess spatial heterogeneity in disease outcomes. First, we present infection curves

for illustrative cities, showing the detailed progress of the infection and its difference from what an SIR model would posit. We also present choropleth maps showing spatial variation in peak infection times, as well as the correlations between the infection trajectory within local areal units and the aggregate infection trajectory for the city as a whole. While an SIR model would predict an absence of systematic variation in the infection curves or the peak infection day for different areal units in the same city, geographically realistic models show considerable disparities in infection progress from one neighborhood to another. To quantify the degree of heterogeneity more broadly, we examine spatial variation in outcomes for each of our city networks. We show that large variations in peak infection days across tracts are typical (often spanning weeks or even months), and that overall correlations of within-tract infection trajectories with the aggregate urban trajectory are generally modest (a substantial departure from what would be expected from an SIR model).

In addition to these relatively abstract metrics, we also examine a simple measure of the potential load on the health care system in each city. Given the locations of each hospital in each city, we attribute infections to each hospital using a Voronoi tessellation (i.e., under the simple model that individuals are most likely to be taken to the nearest hospital if they become seriously ill). Examination of the potential hospital demand over time shows substantial differences in load, with some hospitals severely impacted while others have few cases. Finally, we consider the *social exposure* of individuals to COVID-19, by computing the fraction of individuals with a personal contact who is respectively morbid or deceased. Our model shows considerable differences in these metrics over time, revealing that the pandemic can appear very different to those "on the ground"—evaluating its progress by its impact on their own personal contacts—than what would be suggested by aggregate statistics.

Results

Smooth Aggregate Infection Trajectories Can Mask Local Outbreak Dynamics. When taken over even moderately sized regions, aggregate infection curves can appear relatively smooth. Although this suggests homogeneous mixing (as assumed, e.g., by standard SIR models), appearances can be deceiving. Fig. 1 shows typical realizations of infection curves for two cities (Seattle, WA, and Washington, DC), showing both the aggregate trajectory (red) and trajectories within individual Census tracts (black). While the infection curves in both cases are relatively

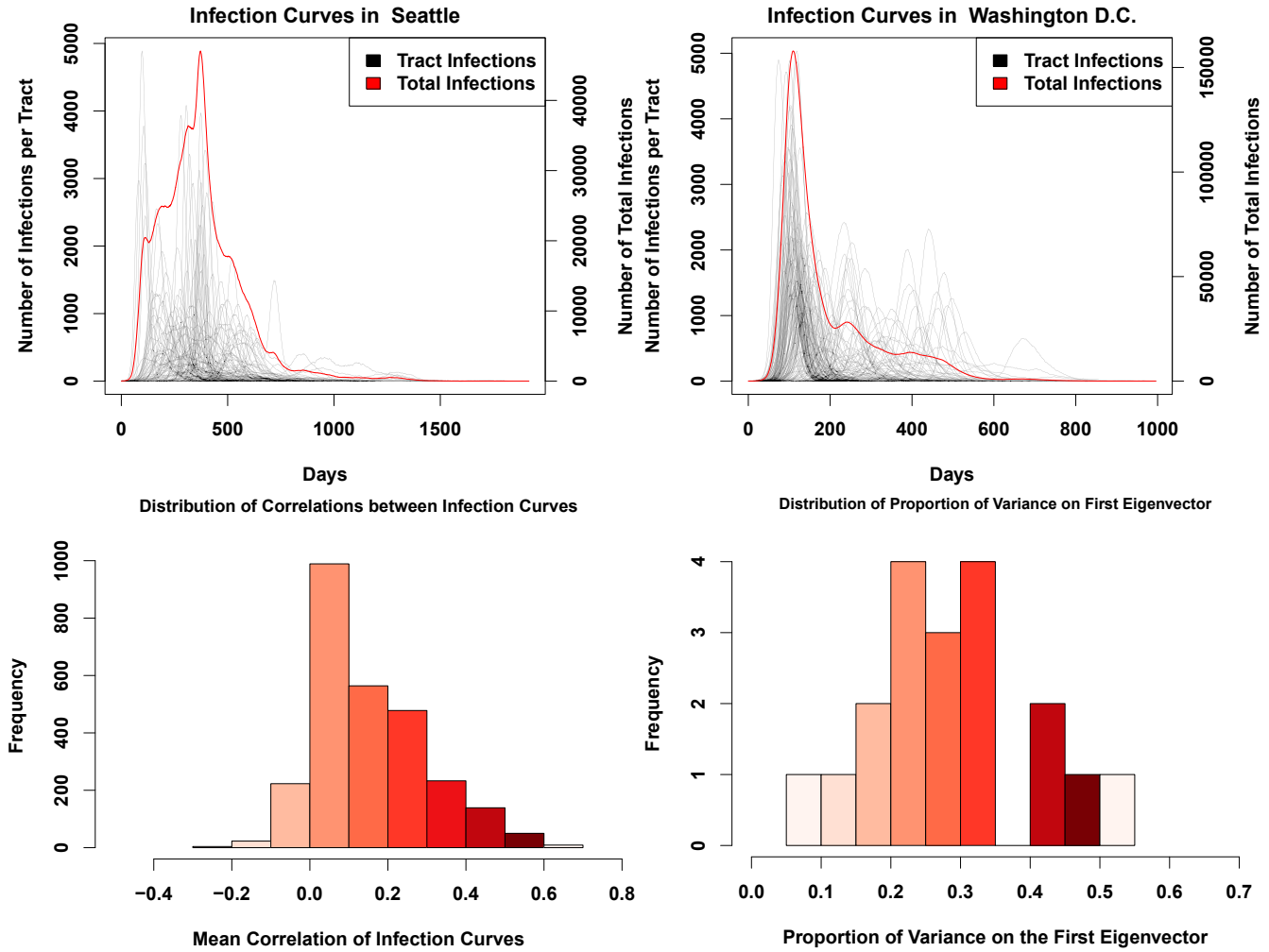


Fig. 1. (*Top Left*) Infection curves for Seattle, WA. The red line is the curve for the whole city, while the black lines are the infection curves for each tract in the city. While the red curve is relatively smooth, this smoothness hides a significant amount of heterogeneity in the timing of the infection curves for each census tract. (*Top Right*) Infection curves for Washington, DC. As with Seattle, the city-level curve conceals considerable spatial variability in the infection's progress. (*Bottom Left*) Histogram showing the mean pairwise correlation of infection curves for each tract within each city, across our entire sample. The infection curve in any given tract is likely to have a correlation of only around 0.2 with any other tract in the city. This histogram includes a single data point for each tract in the sample. (*Bottom Right*) Histogram of variance accounted for by the principal component of the standardized tract-level curve set. None of the principal components account for more than 60% of the variance, with most accounting for around only 35% of the total variance. The data points included here include a single amount of variance explained for each city.

smooth, and suggestive of a fairly simple process involving a sharp early onset followed by an initially sharp but mildly slowing decline in infections, within-tract trajectories tell a different story. Instead of one common curve, we see that tracts vary wildly in onset time and curve width, with some tracts showing peaks weeks or months after the initial aggregate spike has passed.

The cases of Fig. 1 are emblematic of a more systematic phenomenon: The progress of the infection within any given areal unit often has relatively little relationship to its progress in the city as a whole. Fig. 1, *Bottom* assesses this phenomenon over our entire sample, using two different consensus metrics. First, we simply compute the correlation between the infection curves in each pair of tracts (assessed at daily resolution), taking the mean for each tract of its correlation with all other tracts within the city; if the progress of the infection were uniform across the city, the mean correlations would be large and positive. Second, we provide a more direct assessment of the extent to which the set of infection curves can be summarized by a common pattern by taking the variance on the first principal component of the correlation matrix generated from the tract-level correlations discussed immediately above. As before, where different parts of the city experience similar patterns of growth and decline in infections, we expect the dimension of greatest shared variance to account for the overwhelming majority of variation in infection rates. Contrary to these expectations, however, Fig. 1 shows that there is little coherence in tract-level infection patterns. Mean correlations of local infection curves across tracts typically range from ∼0 to 0.5, with a mean of approximately 0.2, indicating very little correspondence between infection timing in one tract and that of another. The principal component analysis tells a similar story: Overall, we see that the first component accounts for relatively little of the total variance in trajectories, with, on average, only around 35% of variation in infection curves lying on the first principal component (and no observed case of the first component accounting for more than 60% of the variance). Interestingly, this variation is not explained by time required for the diffusion process to reach each tract (*SI Appendix*, Fig. S4), in contrast to the hypothesized importance of similar delays in a cross-national context (9). This confirms that local infection curves are consistently distinct, with behavior that is only weakly related to infections in the city as a whole. This is a substantially different scenario than what is commonly assumed in traditional SIR models.

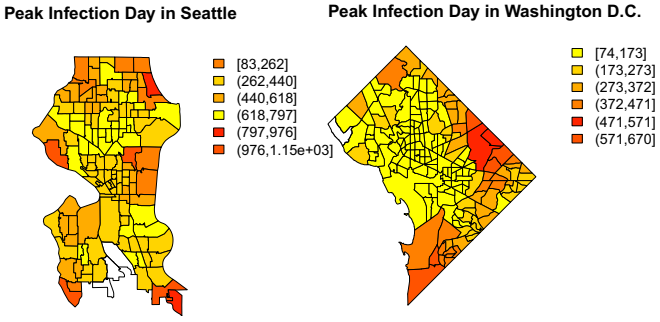


Fig. 2. (*Left*) Choropleth showing the peak day of infection in each tract in the city of Seattle. The map shows significant variability in peak times, with nearby regions sometimes having sharply different patterns. In outlying parts of Seattle, the infection does not peak until almost a year past the first infections, while, in the more eastern and central parts of the city, the infection peaks much earlier. (*Right*) Times to peak infection for Washington, DC tracts. The southern and eastern part of the city has infections that are more delayed than in the central and northern parts of the city. Both of these maps show that there is a high degree of spatial heterogeneity present in the infection curves.

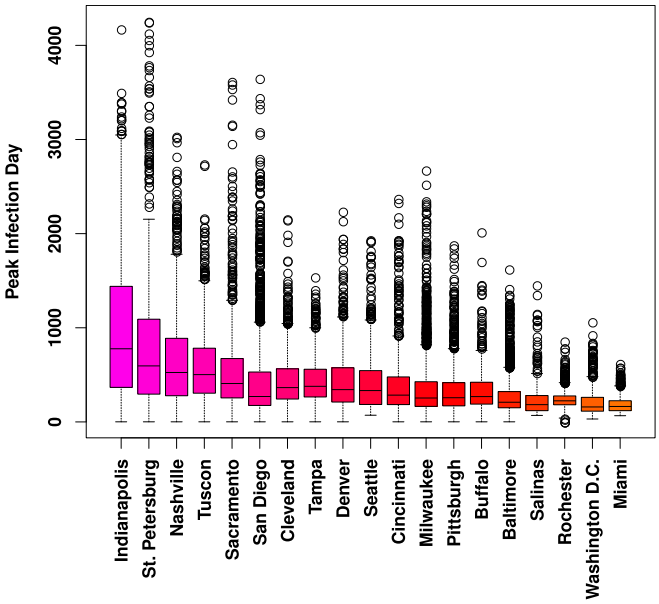


Fig. 3. Marginal distributions of days to peak infection by tract, across 10 simulated trajectories. Although locales vary both in terms of overall median peak time and range of tract-level variation, large differences in peak time are nearly ubiquitous. (Trajectory specific distributions are shown in Fig. S2.)

Peak Infection Days Can Vary Substantially, Even among Nearby Regions. These differences in local infection curves are a consequence of the unevenness of the "social fabric" that spans the city: While the disease can spread rapidly within regions of high local connectivity, it can easily become stalled upon reaching the boundaries of these regions. Further transmission requires that a successful infection event occur via a bridging tie, an event with a potentially long waiting time. Such delays create potential opportunities for public health interventions (trace/isolate/treat strategies), but they can also create a false sense of security for those on the opposite side of the bridge (who may incorrectly assume that their area was passed over by the infection). Indeed, examining the time to peak infection across the cities of Seattle and Washington, DC (Fig. 2) shows that, while peak times are visibly autocorrelated, tracts with different peak times frequently border each other. Residents on opposite sides of the divide may be exposed to very different local infection curves, making risk assessment difficult.

The cases of Seattle and Washington, DC are not anomalous. Looking across multiple trajectories over our entire sample, Fig. 3 shows consistently high variation in per-tract peak infection times for nearly all study communities. (This variation is also seen within individual trajectories, as shown in *SI Appendix*, Fig. S2.) Although peak times in some cities are concentrated within an interval of several days to a week, it is more common for peak times to vary by several months or longer. Extreme delays in peak time arise from "slow burn" trajectories associated with very long infection chains, in which sequential transmission to only one or two alters at a time can sustain infection for greatly extended periods. Such gaps can arise naturally in inhomogeneous networks, but are far from what would be expected under uniform local mixing.

Heterogeneous Impact Timing May Affect Hospital Load. Variation in the timing of COVID-19 impacts across the urban landscape has potential ramifications for health care delivery, creating unequally distributed loads that overburden some providers while leaving others with excess resources. To obtain a sense of how spatial heterogeneity in the infection curve could potentially

impact hospitals, we employ a simple "catchment" model in which seriously ill patients are taken to the nearest hospital, subsequently recovering and/or dying as assumed throughout our modeling framework. Based on prior estimates (36), we assume that 14% of all infections are severe enough to require hospitalization (robustness to alternative rate estimates is shown in *SI Appendix*). While hospitals draw from (and hence average across) areas that are larger than tracts, the heterogeneity shown in Fig. 1 suggests the potential for substantial differences in hospital load over time. Indeed, our models suggest that such differences will occur. Fig. 4 shows the number of patients arriving at each hospital in Seattle and Washington, DC during a typical simulation trajectory. While some hospitals do have demand curves that mirror the city's overall infection curve, others show very different patterns of demand. In particular, some hospitals experience relatively little demand in the early months of the pandemic, only to be hit hard when infections in the city as a whole are winding down.

Just as hospital load varies, hospital capacities vary as well. As a simple measure of strain on hospital resources, we consider the difference between the number of COVID-19 hospitalizations and the total capacity of the hospital (in beds), truncating at zero when demand outstrips supply. (For ease of interpretation as a measure of strain, we take the difference such that higher values indicate fewer available beds.) Using data on hospital locations and capacities, we show, in Fig. 4, strain on all hospitals in Seattle and Washington, DC during a typical infection trajectory. While some hospitals are hardest hit early on (as would be expected from the aggregate infection curve), others do not peak for several months. Likewise, hospitals proximate to areas of the city with very different infection trajectories experience natural "curve flattening," with a more distributed load, while those that happen to draw from positively correlated areas experience very sharp increases and declines in demand. These conditions in some cases combine to keep hospitals well under capacity for the duration of the pandemic, while others are overloaded for long stretches of time. These marked differences in strain for hospitals within the same city highlight the potentially complex consequences of heterogeneous diffusion for health care providers.

Looking across cities, we see the same high-variability patterns as observed in Seattle and Washington. In particular, we note that local variation in disease timing leads to a heavy-tailed distribution for the duration at which hospitals will be at capacity. Fig. 4 shows the marginal distribution of hospital overload periods (defined as total number of days at capacity during the pandemic), over the entire sample. While the most common outcome is for hospitals to be stressed for a brief period (not always to the breaking point), a significant fraction of hospitals end up being overloaded for months—or even, in a small fraction of cases, nearly the whole duration of the pandemic.

It should be reiterated that the hospital load model used here is extremely simplified, and that we are employing a

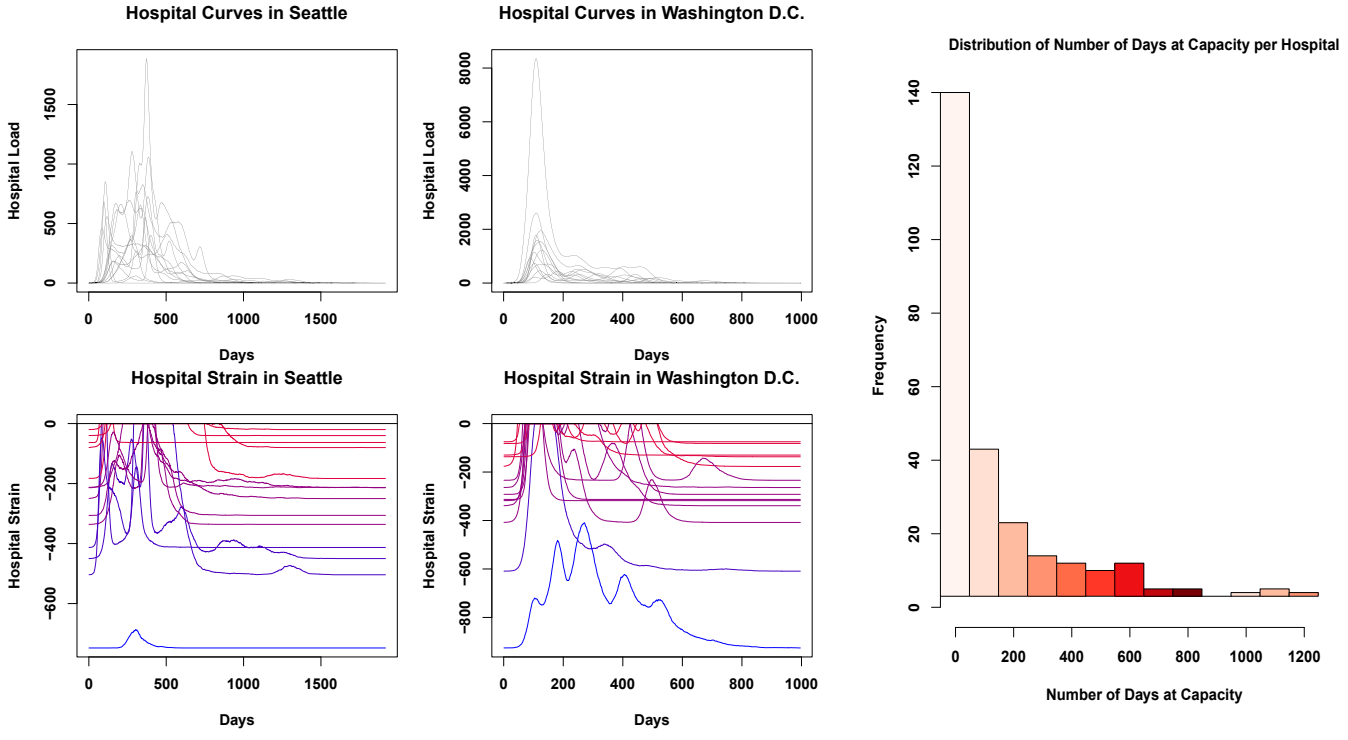


Fig. 4. (*Top Left*) Numbers of infections attributed to each hospital in the city of Seattle, with each curve representing a different hospital. Hospital peak demand times vary markedly, with some getting the majority of their hospitalizations before day 100, and others peaking almost a year into the pandemic. (*Top Middle*) Hospitalizations in Washington, DC. As in Seattle, each hospital has a unique demand trajectory, with some hospitals not getting their peak of infections until more than a year after the infection begins. (*Bottom Left*) Hospital strain in Seattle, WA. Values closer to zero indicate that hospitals are more strained and have fewer open beds, while lower values suggest more resources are available; color varies from blue (low average strain) to red (high average strain). Much like the number of infections, there is a high degree of heterogeneity present here, with hospitals freeing up resources at different points across the first year of the pandemic. (*Bottom Middle*) Hospital strain for Washington, DC. Most hospitals get overwhelmed in the first 25 days of the pandemic, but then are able to recover at different times, usually within the second hundred days of the pandemic; some, however, are hit hard by a second wave, and others remain overwhelmed for several months. (*Right*) Marginal distribution of number of days without available beds, for all hospitals in our sample. While most hospitals will have only brief periods of overload, some will be at or over capacity for the entire pandemic, potentially several years.

no-mitigation scenario. However, these results quite graphically demonstrate that the importance of curve-flattening interventions does not abate once geographical factors are taken into account. On the other hand, these results suggest that differences in hospital load may be substantially more profound than would be anticipated from uniform mixing models, creating logistical challenges and possibly exacerbating existing differences in resource levels across hospitals. At the same time, such heterogeneity implies that resource sharing and patient transfer arrangements could prove more effective as load management strategies than would be suggested by spatially homogeneous models, as hospitals are predicted to vary considerably in the timing of patient demand.

Social Exposures to Morbidity and Mortality Vary by Location. In addition to health care strain, the *subjective experience* of the pandemic will potentially differ for individuals residing in different locations. In particular, social exposures to outcomes such as morbidity or mortality may shape individuals' understandings of the risks posed by COVID-19, and their willingness to undertake protective actions to combat infection. Such exposures may furthermore act as stressors, with potential implications for physical and/or mental health. As a simple measure of social exposure, we consider the question of whether a focal individual (ego) either has experienced a negative outcome themself or has at least one personal contact (alter) who has experienced the outcome in question. (Given the highly salient nature of COVID-19 morbidity and mortality, we focus on the transition to first exposure rather than, e.g., the total number of such exposures, as the first exposure is likely to have the greatest impact on ego's assessment of the potential severity of the disease.)

To examine how social exposure varies by location, we compute the fraction of individuals in each tract who are socially exposed to morbidity or mortality. Fig. 5 shows these proportions for Baltimore, MD, over the course of the pandemic. As with other outcomes examined here, we see considerable variation in timing, with many tracts seeing a rapid increase in exposure to infections, while others go for weeks or months with relatively few persons having a personal contact with the disease. Another notable axis of variation is sharpness. In many tracts, the fraction of individuals with at least one morbid contact transitions from near zero to near one within a matter of days, creating an extremely sharp social transition between the "preexposure world" (in which almost no one present knows someone with the illness) to a "postexposure world" in which almost everyone knows someone with the illness). By contrast, other tracts show a much more gradual increase (sometimes punctuated by jumps), as more and more individuals come to know someone with the disease. In a few tracts that are never hit hard by the pandemic, few people ever have an infected alter; residents of these areas obviously have a very different experience than those of high-prevalence tracts. These distinctions are even more stark for mortality, which takes longer to manifest and which does so much more unevenly. Tracts vary greatly in the fraction of individuals who ultimately lose a personal contact to the disease, and in the rapidity with which that fraction is reached. In many cases, it may take a year or more for this quantity to be realized; until that point, many residents may be skeptical to the notion that the pandemic poses a great risk to them personally.

By way of assessing the milieu within each tract, it is useful to consider the "cross-over" point at which at least half of the residents who will be socially exposed in a given tract have been socially exposed to either COVID-19 morbidity or mortality. Fig. 6 maps these values for Baltimore, MD. It is immediately apparent that social exposures are more strongly spatially autocorrelated than other outcomes considered here, due to the presence of long-range ties within individuals' personal networks. Even so, however, we see strong spatial differentiation, with residents in the urban core being exposed to both morbidity and mortality much more quickly than those on the periphery. This suggests that the social experience of the pandemic will be quite different for those in city centers compared to those in more outlying areas, with the latter taking far longer to be exposed to serious consequences of COVID-19. This may manifest in differences in willingness to adopt protective actions, with those in the urban core being more highly motivated to take action (and perhaps resistant to rhetoric downplaying the severity of the disease) than those on the outskirts of the city.

Discussion

Our simulation results all underscore the potential effects of local spatial heterogeneity on disease spread. The spatial heterogeneity driving these results occurs on a very small scale (i.e., Census blocks), operating well below the level of the city a whole. As the infection spreads, relatively small differences in local network connectivity and the prevalence of bridging ties driven by uneven population distribution can lead to substantial differences in infection timing and severity, leading different areas in each city to have vastly different experiences of the pandemic. Resources will be utilized differently in different areas, as some areas will experience the bulk of their infections far later than others, and the subjective experience of a given individual regarding the pandemic threat may differ substantially from someone in another area. These behaviors are in striking contrast to what is assumed by models based on the assumption of

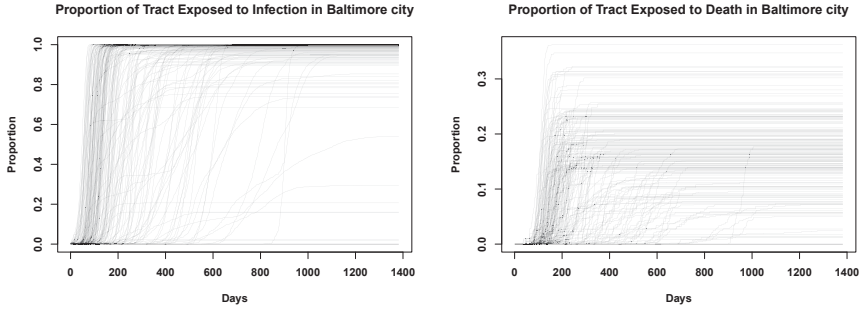


Fig. 5. (*Left*) Trajectories showing the fraction of people in each tract in Baltimore who have an infected person in their personal network across time. We see a large degree of spatial heterogeneity, as some tracts are more insulated from others in terms of social exposure. However, by the end of the pandemic, most people across all tracts have been exposed to someone who has had the disease. (*Right*) The fraction of persons in each tract who have an alter who died from COVID-19 in their personal network. On average, only around 20 to 30% of people in any given tract know someone who died, by the end of the pandemic, although this varies widely across tracts.

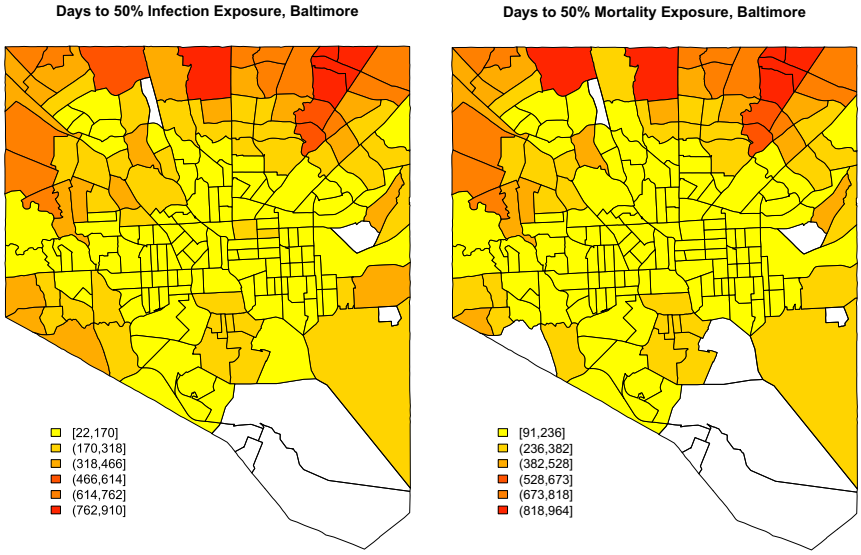


Fig. 6. (*Left*) Choropleth showing the time for half of those in each tract to be socially exposed to COVID-19 morbidity in Baltimore, MD. The central parts of the city are exposed far sooner than the northwestern part of the city. (*Right*) Choropleth showing the time for half of those in each tract to be socially exposed to COVID-19 mortality. Central Baltimore is exposed to deaths in personal networks far sooner than the more outlying areas of the city.

spatially homogeneous mixing, which posit uniform progress of the infection within local areas.

As noted at the outset, our model is based on a no-mitigation scenario, and is not intended to capture the impact of social distancing. While distancing measures by definition limit transmission rates—and will hence slow diffusion—contacts occurring through spatially correlated networks like those modeled here are still likely to show patterns of heterogeneity like those described. One notable observation from our simulations is the long outbreak delay that some census tracts experience, even in the absence of social distancing. This would suggest that relaxation of mitigation measures leading to a resumption of "normal" diffusion may initially appear to have few negative effects, only to lead to deadly outbreaks weeks or months later. Public health messaging may need to stress that apparent lulls in disease progress are not necessarily indicators that the threat has subsided, and that areas "passed over" by past outbreaks could be impacted at any time.

Finally, we stress that conventional diffusion models using locally homogeneous mixing have been of considerable value in both pandemic planning and scenario evaluation. Our findings should not be taken as an argument against the use of such models. However, the observation that incorporating geographical heterogeneity in contact rates leads to radically different local behavior would seem to suggest that there is value in including such effects in models intended to capture outcomes at the city or county level. Since these are the scales on which decisions regarding infrastructure management, health care logistics, and other policies are often made, improved geographical realism could potentially have a substantial impact on our ability to reduce lives lost to the COVID-19 pandemic.

Data Availability. Social networks and analysis code data have been deposited in Harvard Dataverse (https://doi.org/10.7910/DVN/B9XKSR) (37).

ACKNOWLEDGMENTS. This material is based on research supported by NSF Awards IIS-1939237 and SES-1826589 to C.T.B., and by a University of California, Irvine Council on Research, Computing and Libraries grant.

1. World Health Organization, "Coronavirus disease 2019 (COVID-19): Situation report" (Rep. 205, World Health Organization, 2020).
2. G. Onder, G. Rezza, S. Brusaferro, Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *JAMA* **323**, 1775–1776 (2020).
3. F. Jiang *et al.*, Review of the clinical characteristics of coronavirus disease 2019 (COVID-19). *J. Gen. Intern. Med.* **35**, 1545–1549 (2020).
4. Y. J. Geng *et al.*, Pathophysiological characteristics and therapeutic approaches for pulmonary injury and cardiovascular complications of coronavirus disease 2019. *Cardiovasc. Pathol.* **47**, 107228 (2020).
5. World Health Organization, "Coronavirus disease 2019 (COVID-19): Situation report" (Rep. 43, World Health Organization, 2020).
6. M. L. Jackson *et al.*, Effects of weather-related social distancing on city-scale transmission of respiratory viruses. *medRxiv*:10.1101/2020.03.02.20027599 (3 March 2020).
7. Y. Zhang, B. Jiang, J. Yuan, Y. Tao, The impact of social distancing and epicenter lockdown on the COVID-19 epidemic in mainland China: A data-driven SEIQR model study. *medRxiv*:10.1101/2020.03.04.20031187 (6 March 2020).
8. B. S. Pujari, S. M. Shekatkar, Multi-city modeling of epidemics using spatial networks: Application to 2019-nCov (COVID-19) coronavirus in India. *medRxiv*:10.1101/2020.03.13.20035386 (17 March 2020).
9. D. Brockmann, D. Helbing, The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–1342 (2013).
10. E. S. Spiro, Z. W. Almquist, C. T. Butts, The persistence of division: Geography, institutions, and online friendship ties. *Socius* **2**, 2378023116634340 (2016).
11. E. J. Smith *et al.*, The relationship of age to personal network size, relational multiplexity, and proximity to alters in the Western United States. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **70**, 91–99 (2015).
12. Q. Wang, N. E. Phillips, M. L. Small, R. J. Sampson, Urban mobility and neighborhood isolation in America's 50 largest cities. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 7735–7740 (2018).
13. Z. W. Almquist, C. T. Butts, Predicting regional self-identification from spatial network models. *Geogr. Anal.* **47**, 50–72 (2015).
14. S. Riley, Large-scale spatial-transmission models of infectious disease. *Science* **316**, 1298–1301 (2007).
15. J. R. Hipp, C. T. Butts, R. Acton, N. N. Nagle, A. Boessen, Extrapolative simulation of neighborhood networks based on population spatial distribution: Do they predict crime? *Soc. Network.* **35**, 614–625 (2013).
16. R. J. Sampson, P. Sharkey, Neighborhood selection and the social reproduction of concentrated racial inequality. *Demography* **45**, 1–29 (2008).
17. Z. W. Almquist, C. T. Butts, Point process models for household distributions within small areal units. *Demogr. Res.* **26**, 593–632 (2012).
18. D. S. Massey, N. A. Denton, *American Apartheid: Segregation and the Making of the Underclass* (Harvard University Press, 1993).
19. D. S. Massey, J. Tannen, Suburbanization and segregation in the United States: 1970–2010. *Ethn. Racial Stud.* **41**, 1594–1611 (2018).
20. M. D. Verhagen, D. M. Brazel, J. B. Dowd, I. Kashnitsky, M. Mills, Mapping hospital demand: Demographics, spatial variation, and the risk of "hospital deserts" during COVID-19 in England and Wales. *OSF*:10.31219/osf.io/g8s96 (21 March 2020).
21. J. B. Dowd *et al.*, Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9696–9698 (2020).
22. P. Block *et al.*, Social network-based distancing strategies to flatten the COVID 19 curve in a post-lockdown world. *Nat. Human Behav.* **4**, 588–596 (2020).

23. C. T. Butts, R. M. Acton, J. R. Hipp, N. N. Nagle, Geographical variability and network structure. *Soc. Network*. **34**, 82–100 (2012).
24. Department of Homeland Security, Homeland infrastructure foundation-level data (HIFLD). https://hifld-geoplatform.opendata.arcgis.com/datasets/hospitals. Accessed 3 April 2020.
25. New York Times, Coronavirus (COVID-19) data in the United States. https://github.com/nytimes/covid-19-data. Accessed 5 May 2020.
26. R Core Team, *R: A Language and Environment for Statistical Computing*, (R Foundation for Statistical Computing, Vienna, Austria, 2020).
27. C. T. Butts, network: A package for managing relational data in R. *J. Stat. Software* **24.2**, 1–36 (2008).
28. C. T. Butts, Social network analysis with sna. *J. Stat. Software* **24.6**, 1–51 (2008).
29. M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, M. Morris, statnet: Software tools for the representation, visualization, analysis and simulation of network data. *J. Stat. Software* **24**, 1–11 (2008).
30. M. Morris, *Network Epidemiology: A Handbook for Survey Design and Data Collection* (Oxford University Press, 2004).
31. C. Viboud *et al.*, Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447–451 (2006).
32. L. Li *et al.*, Analysis of transmission dynamics for Zika virus on networks. *Appl. Math. Comput.* **347**, 566–577 (2019).
33. C. T. Butts, R. M. Acton, "Spatial modeling of social networks" in *The Sage Handbook of GIS and Society Research*, T. L. Nyerges, Ed. (SAGE, Thousand Oaks, CA, 2011), pp. 222–250.
34. Z. W. Almquist, Large-scale spatial network models: An application to modeling information diffusion through the homeless population of San Francisco. *Environ. Plann. B* **47**, 523–540 (2020).
35. G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Ravenio, 2016).
36. E. K. Stokes *et al.*, Coronavirus disease 2019 case surveillance—United States, January 22–May 30, 2020. *MMWR (Morb. Mortal. Wkly. Rep.)* **69**, 759–765 (2020).
37. L. J. Thomas *et al.*, Replication data for: Spatial heterogeneity can lead to substantial local variations in COVID-19 timing and severity. Harvard Dataverse. https://doi.org/10.7910/DVN/B9XKSR. Deposited 26 August 2020.

Social Networks 76 (2024) 51–67

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Parameter estimation procedures for exponential-family random graph models on count-valued networks: A comparative simulation study

Peng Huang [a], Carter T. Butts [a,b,*]

[a] Departments of Sociology, and Statistics, University of California, Irvine, United States of America.
[b] Departments of Computer Science, and EECS, University of California, Irvine, United States of America

ARTICLE INFO

Keywords:
Contrastive divergence
Exponential-family random graph model
Markov chain Monte Carlo
Maximum likelihood estimation
Pseudo likelihood
Valued/Weighted networks

ABSTRACT

The exponential-family random graph models (ERGMs) have emerged as an important framework for modeling social networks for a wide variety of relational types. ERGMs for valued networks are less well-developed than their unvalued counterparts, and pose particular computational challenges. Network data with edge values on the non-negative integers (count-valued networks) is an important such case, with examples ranging from the magnitude of migration and trade flows between places to the frequency of interactions and encounters between individuals. Here, we propose an efficient parallelizable subsampled maximum pseudo-likelihood estimation (MPLE) scheme for count-valued ERGMs, and compare its performance with existing Contrastive Divergence (CD) and Monte Carlo Maximum Likelihood Estimation (MCMLE) approaches via a simulation study based on migration flow networks in two U.S. states. Our results suggest that edge value variance is a key factor in method performance, while network size mainly influences their relative merits in computational time. For small-variance networks, all methods perform well in point estimations while CD greatly overestimates uncertainties, and MPLE underestimates them for dependence terms; all methods have fast estimation for small networks, but CD and subsampled multi-core MPLE provides speed advantages as network size increases. For large-variance networks, both MPLE and MCMLE offer high-quality estimates of coefficients and their uncertainty, but MPLE is significantly faster than MCMLE; MPLE is also a better seeding method for MCMLE than CD, as the latter makes MCMLE more prone to convergence failure. The study suggests that MCMLE and MPLE should be the default approach to estimate ERGMs for small-variance and large-variance valued networks, respectively. We also offer further suggestions regarding choice of computational method for valued ERGMs based on data structure, available computational resources and analytical goals.

1. Introduction

Binary relations - relations in which edges can be approximated as simply "present" or "absent" - form the backbone of the social network field, with decades of theoretical, methodological, and empirical progress in understanding their structure and function. Valued relations, while by no means neglected, are less well-understood, and our tools for studying them less well-developed. Yet, the "strength of social ties" is at core of many scientific questions in a range of social settings (Granovetter, 1973; McMillan, 2022). Examples of valued relations include the frequency of interaction in interpersonal contact networks (Bernard et al., 1979), number of cosponsored bills shared among legislators (Cranmer and Desmarais, 2011; Fowler, 2006), communication volume within and among organizations (Drabek et al., 1981; Butts et al., 2007), encounters among non-human animals (Faust, 2011), and trade and migration flows among nations (Windzio, 2018; Ward et al., 2013). The need for rich information about social relations is particularly acute for networks involving interactions among aggregate entities such as nations, geographical areas, gangs, or formal organizations: because ties in such networks are themselves frequently aggregations of lower-level interactions, it is often the case that one's interest is not in the mere existence of trade, migration, homicide, communication, or other interactions, but their volume, frequency, or other quantitative features. In such settings, modeling edge states is of considerable substantive importance.

The earliest statistical modeling of valued relations was accomplished via network regression methods (Krackhardt, 1988); these provide only least-squares estimates of covariate effects, although autocorrelation-robust null hypothesis tests for such effects are well-known (Dekker et al., 2007), and some generalization via generalized linear models (GLMs) and related techniques is possible. Some forms

* Correspondence to: Departments of Sociology, Statistics, Computer Science, and EECS, University of California-Irvine, SSPA 2145, Irvine, CA 92697, USA.
E-mail address: buttsc@uci.edu (C.T. Butts).

https://doi.org/10.1016/j.socnet.2023.07.001

of dependence can, further, be controlled semi-parametrically using latent structure models (e.g., Nowicki and Snijders, 2001; Hoff et al., 2002; Vu et al., 2013; Aicher et al., 2014), allowing estimation of covariate effects while accounting for unobserved mechanisms that can be written in terms of mixing on unobserved variables. Parametric models for valued graphs with general classes of dependence effects have been longer in coming, the current state of the art being exponential family random graph models (ERGMs) defined on sets of valued graphs (Block et al., 2022; Desmarais and Cranmer, 2012a; Krivitsky, 2012; Krivitsky and Butts, 2017); but see also Robins et al. (1999) for a pioneering example using categorical data and pseudo-likelihood estimation. Although ERGMs for valued graphs are not complete in the sense that they are for unvalued graphs (i.e., for most types of edge values, it is not always possible to write an arbitrary distribution on the order-N valued graphs in ERGM form), they are still highly general families, able to flexibly specify a wide range of effects. Since their introduction, they have been applied in a number of settings, ranging from networks of collaboration in government, and networks of migration flows, to networks of functional connectivity between brain regions (Huang and Butts, 2022; Simpson et al., 2013; Ulibarri and Scott, 2017; Windzio, 2018).

Notwithstanding their broad applicability, parameter estimation for ERGMs in practice can be computationally demanding, a problem that is especially acute for valued networks. This issue has clearly had an impact on empirical network analyses in the published literature, forcing researchers to employ compromises or workarounds. As an example, Aksoy and Yıldırım (forthcoming) noted in their paper that they could not obtain convergence for a single 81-node network using valued ERGMs. For research that managed to obtain ERGM estimation of their valued networks, they had to either dichotomize the data and fall back to binary models (Leal, 2021), or coarsen the counts into quintiles (Windzio, 2018; Windzio et al., 2019); data transformation of this type greatly reduces computational difficulties, but in the meantime brings information loss and underestimation of variability (Altman and Royston, 2006). In short, even though methodological advances in valued network modeling have made it possible for researchers to capture quantitative features of relations beyond dichotomizational operations (Cranmer and Desmarais, 2011), the computational load remains a lingering hurdle to fully exploit the potential of these methods in scientific applications.

The major computational cost of ERGM estimation comes from the normalizing factor in its likelihood function, which is generally an intractable function involving the sum or integral of an exponentiated potential over the set of all possible network configurations. Although much is made over the fact that these sums have too many elements to explicitly evaluate (except in the case of extremely small unvalued graphs, e.g. Vega Yon et al., 2021), this is not the major obstacle to computation: rather, the difficulty rises from the extreme roughness (i.e., high variance) of the exponentiated potential over the support, which (in the absence of an explicit analytical solution) renders naive attempts at numerical approximation ineffective. This problem can be amplified in the valued case, particularly where edge values vary greatly; valued edges can also pose challenges for some approximate estimation procedures that are successful in the case of unvalued ERGMs, as they must now explore a larger *per edge* state space. This high cost of estimation puts a priority on computationally efficient approximation methods. However, there has not been to date a systematic study of how well such methods perform, either in terms of improved computational efficiency or quality of estimation.

This paper provides a look at this issue, evaluating estimation quality and computational cost for a number of alternative valued ERGM estimation techniques. We focus on ERGMs for count-valued networks, i.e. relations whose edges take values on the unbounded non-negative integers, evaluating estimators via a simulation study based on intercounty migration-flow networks in two U.S. states. We vary the variance of edge values and the (node) size of the network, to simulate different data structures. The methods examined include the two currently implemented "standard" strategies - contrastive divergence (CD; Krivitsky, 2017) and Markov Chain Monte Carlo maximum likelihood estimation (MCMLE; Hunter et al., 2012) - as well as one approach not previously used in this setting, maximum pseudo-likelihood estimation (MPLE). MPLE is a workhorse approximation method in the binary ERGM case (Strauss and Ikeda, 1990), but requires special implementation measures for the count-data case, and to our knowledge has not previously been used for count-data ERGMs with general dependence. We also compare the performance of MPLE and CD as two seeding options for MCMLE. For all methods, we evaluate their computational speed, bias, variability, accuracy, calibration of estimated standard errors and confidence coverage.

The remainder of the paper proceeds as follows. Section 2 briefly reviews ERGMs for valued networks, with applicable estimation strategies discussed in Section 3. Our simulation study design is described in Section 4, with results reported in Section 5. Section 6 discusses implications for method selection, and Section 7 concludes the paper.

2. Count-valued ERGMs

An ERGM family for count data can be written as

$$\Pr(Y = y|\theta, X) = \frac{h(y)\exp(\theta^T g(y, X))}{\sum_{y'\in\mathscr{Y}} h(y')\exp(\theta^T g(y', X))}, \tag{1}$$

where y is a realization of the network random variable Y on support $\mathscr{Y}$, the elements of which are graphs whose edges take values on the set $\{0, 1, \ldots\}$. (Here, we further assume that $\mathscr{Y}$ is a subset of the order-N count-valued graphs, though generalization is possible.) $g : \mathscr{Y}, X \mapsto \mathbb{R}^k$ is a vector of sufficient statistics, determined by exogenous covariates X and the graph state y, with corresponding parameter vector θ. Finally, $h : \mathscr{Y} \mapsto \mathbb{R}_{\geq 0}$ is the *reference measure*, which defines the limiting behavior of the model as $\theta \to 0$. Often tacitly taken to be constant for binary ERGMs, the reference measure is essential for valued ERGMs, as it determines the marginal distribution of edge values under the reference (Krivitsky, 2012). Leaving $h(y) \propto 1$ leads to a marginal Boltzmann baseline distribution, while choosing

$$h(y) = \prod_{(i,j)\in\mathbb{Y}} (y_{ij}!)^{-1} \tag{2}$$

where $\mathbb{Y}$ is the set of edge variables, and y_{ij} is the value of the (i, j) edge, leads to a marginal Poisson baseline distribution of edge values. Other choices are also possible, some of which may have specific substantive interpretations (see e.g. Butts, 2019, 2020, for examples in the binary case).

As with binary ERGMs, we may specify the conditional probability that a given i, j edge variable will take a specified value. Again interpreting Y and y as random adjacency matrices, let Y^c_{ij} (respectively y^c_{ij}) refer to the set of all edge variables other than the ijth, and let the notation $z \cup Y^c_{ij}$ refer to the network formed by Y with the ijth edge variable set to value z. Then we have

$$\Pr(Y_{ij} = y_{ij}|Y^c_{ij} = y^c_{ij}, \theta, X) = \frac{h(y_{ij} \cup y^c_{ij})\exp(\theta^T g(y_{ij} \cup y^c_{ij}, X))}{\sum_{\ell=0}^{\infty} h(\ell \cup y^c_{ij})\exp(\theta^T g(\ell \cup y^c_{ij}, X))} \tag{3}$$

$$= \left[\sum_{\ell=0}^{\infty} \frac{h(\ell \cup y^c_{ij})}{h(y_{ij} \cup y^c_{ij})} \exp\left[\theta^T\left(g(\ell \cup y^c_{ij}, X) - g(y_{ij} \cup y^c_{ij}, X)\right)\right]\right]^{-1}. \tag{4}$$

While the derivation is identical to the binary case (as can be appreciated by noting that Eq. (3) would reduce to the usual logistic form if ℓ were restricted to be ≤ 1), we note the computationally important difference that the conditional edge probability itself now has a non-trivial normalizing factor. In the general case, this has no analytical solution, and since it involves an infinite sum it cannot be explicitly evaluated otherwise. Although this does not impact e.g. the acceptance calculations for typical Markov Chain Monte Carlo (MCMC) algorithms (since the conditional odds of one graph versus another does

not depend upon either normalizing factor), it does affect computation for the MPLE (which does depend on the conditional edge probability). Here, we formulate a finite sum approximation to Eq. (3) for MPLE, as described below.

3. Estimation strategies for count-valued ERGMs

While many approaches to parameter estimation are possible, we focus here on approximations to the maximum likelihood estimator (MLE). Here, we briefly review the strategies employed, including special considerations for the count-data case. We note that some alternative schemes explored in the binary case (e.g., variational methods: Mele, 2017; Tan and Friel, 2020; Wainwright et al., 2008) may be adapted to the count data problem, but for purposes of this paper we limit our study to approaches that have been established as broadly useful, and for which count-valued implementations currently exist (with the exception of MPLE, which we extend).

3.1. Monte Carlo maximum likelihood estimation

There are currently two widely used schemes for MCMC-based maximum likelihood estimation: stochastic approximation (Snijders, 2002; Wang et al., 2009), which is based on attempting to match the expected sufficient statistics to their observed values (exploiting the coincidence of methods-of-moments and MLE for exponential families); and the Geyer–Thompson method (Geyer and Thompson, 1992; Hunter et al., 2008b) (supplemented in current implementations by Hummel stepping Hummel et al., 2012), which uses an importance sampling scheme to directly optimize the log-likelihood surface. We here employ the former in its `statnet` implementation (Hunter et al., 2008b; Krivitsky et al., 2012).

MCMLE methods are the current gold-standard techniques for ERGM maximum likelihood estimation, with good theoretical properties (Snijders, 2002; Handcock, 2003) and strong performance in simulation studies for binary networks (van Duijn et al., 2009). An important bottleneck impacting the use of MCMLE, however, is the ability to produce relatively high-quality draws from the specified ERGM distribution (without which, the algorithms will not converge correctly). While it is known that conventional MCMC algorithms can in principle mix arbitrarily slowly (Snijders, 2002; Bhamidi et al., 2011), in actual practice this problem has been observed primarily in badly specified models that are degenerate or near-degenerate, and hence of limited relevance in typical social network applications (Hunter et al., 2012). That said, estimation time can still become long on very large networks, particularly for models with strong edgewise dependence.

This cost issue becomes more acute for valued ERGMs, especially where edge values are highly variable. Intuitively, good MCMC mixing requires the Markov chain to explore the space of high-probability graphs, whose size increases substantially when edge values vary over a large range. For instance, for a simple random walk MCMC algorithm that proposes perturbing edges at random,[1] $\mathcal{O}(N^2)$ toggles may be needed to ensure that every edge variable in an unvalued graph has a high probability of having the "opportunity" to change state. If edges typically vary over some interval of order R, then a similar random walk scheme that increments or decrements edge values will need at least $\mathcal{O}(RN^2)$ for each edge to have the "opportunity" to cover its range of values. For networks with large counts (e.g., migrant-flow networks), one can easily obtain $R \gg N$, in which case simulation costs can rapidly become prohibitive. Although this problem can be alleviated by coarsening edge values to a much smaller range (as done e.g. by Windzio, 2018; Windzio et al., 2019), this both loses information and distorts the resulting model (since e.g., coarsening artificially reduces the entropy of the graph distribution). In principle, improved MCMC algorithms offer a better way to address this problem in the long-term, but current implementations do not seem to scale well for high-variance count models (e.g. Aksoy and Yıldırım, forthcoming). As we show below, MPLE can often deliver comparable estimation quality to MCMLE for high-R valued ERGMs, where the latter suffers substantial increases in computational cost.

3.2. Contrastive divergence

One alternative to either numerical approximation of expected statistics or of log likelihood ratios is to use a local approximation to the gradient of the likelihood in regions of the support "near" the observed data. This is the essential idea behind contrastive divergence (CD) (Hinton, 2002), a method originally introduced in the machine learning literature for scalable inference that is particularly well-suited to ERGMs and other exponential families (Asuncion et al., 2010; Krivitsky, 2017). CD can be employed for both valued and unvalued graphs, and greatly reduces computational time by using only very short MCMC chains starting at the observed data, depending on neither sample convergence nor burn-in. It is, however, an approximate technique that optimizes a function closely related to the pseudo-likelihood (Asuncion et al., 2010), and thus shares some of the drawbacks of the MPLE. These properties make CD a reasonable seeding method that offers MCMLE with starting values for estimators, as starting values close to the MLE is known to help reduce iteration rounds and avoid convergence failures for MCMC algorithms. Krivitsky (2017) found that MPLE typically outperformed CD as a seeding method for MCMLE in the binary ERGM regime; but since MPLE has not yet been implemented for valued ERGMs, CD currently serves as the default seeding method for MCMLE in the `statenet` package for valued graphs. Here, we evaluate CD both as a standalone method and a seeding method, in comparison with MPLE, for MCMLE.

3.3. Maximum pseudo-likelihood estimation

Although maximum pseudo-likelihood estimation (MPLE) has not to our knowledge been studied or implemented for count-valued ERGMs, it is an otherwise well-known technique (being the first practical method for general ERGM estimation Strauss and Ikeda, 1990). MPLE optimizes the product of the conditional likelihoods of each edge variable (the eponymous pseudo-likelihood Besag, 1974). In the unvalued case, this reduces to a logistic regression problem, allowing the MPLE to be obtained using standard regression algorithms (a fact that was instrumental in its early adoption, see e.g. Anderson et al., 1999). The MPLE is known to be consistent in some asymptotic scenarios (Hyvärinen, 2006; Strauss and Ikeda, 1990). For finite scenarios, in the special case of edgewise independent ERGMs, the MPLE coincides with the MLE; this ceases to be true for dependence models, though the MPLE is generally close enough to the MLE to be used as a standard method for initializing MCMLE estimators, and its first-order performance on large networks can be very good (An, 2016; Schmid and Desmarais, 2017). Because it does not fully account for interactions among edge variables, the pseudo-likelihood function tends to be excessively concentrated, leading to poor calibration of standard error estimates (as shown in binary ERGMs: Lubbers and Snijders, 2007; van Duijn et al., 2009). However, MPLE computation can be quite efficient, further aided by the fact that (1) the pseudo-likelihood itself can be approximated by subsampling edge variables, rather than computing on all of them, and (2) the calculations in question are embarrassingly parallel, making it possible to greatly reduce wall-clock time on multi-core CPUs.

[1] Practical implementations often use slightly different proposals, but the basic intuition carries.

As noted above, MPLE computation in the count-data context is more complex than in the binary case, and to our knowledge it has not been previously studied for count-valued ERGMs with dyadic dependence. We thus consider it here in greater detail. As in the binary case, the MPLE is defined by

$$\hat{\theta}_{\text{MPLE}} = \arg\max_{\theta} \prod_{(i,j)\in\mathbb{Y}} \Pr(Y_{ij} = y_{ij} | Y^c_{ij} = y^c_{ij}, \theta, X), \tag{5}$$

where the conditional probabilities in question are given by Eqs. (3) and (4). Per Eq. (4), these latter conditionals depend upon a sum over the possible edge states of products of two factors: one involves the ratio of the reference measure at the observed edge value versus its alternative values, and the other involves the exponentiated difference in sufficient statistics between the observed network and the same network with the focal edge taking on alternative values. For the former, we observe that (in the case of the Poissonian reference), we have

$$\frac{h(\ell \cup y^c_{ij})}{h(y_{ij} \cup y^c_{ij})} = \frac{(\ell!)^{-1} \prod_{(k,l)\in\mathbb{Y}\setminus(i,j)} (y_{kl}!)^{-1}}{(y_{ij}!)^{-1} \prod_{(k,l)\in\mathbb{Y}\setminus(i,j)} (y_{kl}!)^{-1}} = \frac{y_{ij}!}{\ell!}, \tag{6}$$

while the latter is simply

$$\exp\left[\theta^T \Delta_{ij}(y, \ell)\right],$$

where Δ is the "generalized" changescore

$$\Delta_{ij}(y, \ell) = g(\ell \cup y^c_{ij}, X) - g(y_{ij} \cup y^c_{ij}, X).$$

There is not, in general, a simple form for the sum of these terms over all ℓ. However, we observe that the ratio of Eq. (6) falls very rapidly (as roughly $\ell^{-\ell}$) for $\ell \gg y_{ij}$, and it is hence possible in practice to approximate the infinite sum by truncation. More generally, we employ several techniques for improving computational performance, as described in the following subsections.

3.3.1. Pre-caching of ratios and differences

We note that neither the ratio of reference measures nor the changescores depend upon θ. Considerable computational savings can hence be had by pre-computing the ratios of Eq. (6) and the Δ values for the necessary range of ℓ values on each edge. This carries a storage cost that scales with the product of the ℓ range and the number of edge variables used, but avoids frequent recalculation of these (expensive) quantities on each pseudo-likelihood evaluation.

3.3.2. Edge sum truncation and/or coarsening

Per Eq. (3), the conditional log-likelihood of each edge variable involves as sum over $\ell \in 0, \ldots, \infty$. As noted above, we may approximate this sum by instead evaluating it over $\ell \in 0, \ldots, \ell_{\max}$, where $\ell_{\max}$ is large enough to be dominated by the decline in $y_{ij}!/\ell!$. Where the marginal distributions of each edge variable can be approximated as roughly Poissonian, choosing $\ell_{\max} = \lambda \max_{(i,j)\in\mathbb{Y}} y_{ij}$ with e.g. $\lambda \approx 4$ is an extremely conservative approach. (This is based on the observation that a Poisson random variable with expectation $z \geq 2$ has a 99.9% quantile for $4z$, assuming conservatively the expectation is the maximum observed value). λ can be further reduced as its max observed value increases (because the quantile of λz grows with z).

Truncation using the above method is adequate for small networks, or networks with low edge counts. However, when edge counts become extremely large, considerable computational effort may be wasted in computing conditional probabilities for small ℓ values when the observed value is large, or for large ℓ when the observed value is small. Using the same Poissonian approximation, we may further improve performance by working with the edgewise doubly-truncated sum over $\ell \in \ell^{ij}_{\min}, \ldots, \ell^{ij}_{\max}$, with $\ell^{ij}_{\min} = \max[0, y_{ij} - 4\lambda\sqrt{y_{ij}}]$ and $\ell^{ij}_{\max} = y_{ij} + 4\lambda\sqrt{y_{ij}}$. Because it is common to have network effects that can strongly suppress edges, however, we also recommend retaining some very small edge values as a buffer. Valued social networks usually have right-skewed distributions of edge values, so adding a few small edge values can also effectively cover the empirical distribution without significant increase in computation load. Our code by default retains integers from 0 to 5, although we strongly encourage extending coverage to the closest integer of sample mean of y, $[\mu]$ when feasible. This also defines the support of edges whose observed value is zero. The approach then leads to sums over ℓ values of the form $\ell \in \{0, \ldots, [\mu]\} \cup \{\ell^{ij}_{\min}, \ldots, \ell^{ij}_{\max}\}$. This usually retains $\mathcal{O}(\sqrt{y_{ij}})$ terms per sampled edge, which is often a substantial savings as y_{ij} values become large.

When dealing with extremely large counts, storing and computing even $\sqrt{y}$ terms can become prohibitive (particularly if many edge variables are needed for adequate statistical power). In such cases, a coarsened approximation to the sum is another option. To coarsen, we select k evenly spaced values from $\ell^{ij}_{\min}, \ldots, \ell^{ij}_{\max}$, and compute the associated contributions to the edge sum only for these terms. We also include, however, the gap (in terms of the number of "skipped" ℓ values) between subsequent calculated terms, and weight each computed term by the number of elements in the gap; this is equivalent to approximating the sum via a step function, with knots at the computed ℓ values. Our experience with this method has been promising, although problems can ensue if the sum becomes heavily concentrated on a range of terms that lie within adjacent knots. We thus do not employ this technique in this paper, although we offer it as a promising target for future research. Of course, other approximation methods are also possible (e.g., integral approximations), and may be useful in the large-count regime.

3.3.3. Edge variable sampling

Although the exact calculation of the pseudo-likelihood is at least $\mathcal{O}(RN^2)$, the log pseudo-likelihood can easily be approximated by random sampling of edge variables; this reduces both storage and computational cost. As shown by our experiments, subsampled MPLE can yield high-quality estimates with less time consumed. Our implementation offers different sampling schemes such as uniform random sampling, as well as weighted (i.e., importance) sampling schemes analogous to the "Tie-No Tie" proposal method frequently used in ERGM MCMC (Morris et al., 2008). The two schemes are almost identical in our study case because the binary density is close to 0.5, and we use the random sampling scheme in this paper for simplicity.

3.3.4. Parallel evaluation of conditional log-likelihoods

Because the log of the pseudo-likelihood is linearly separable, its calculation is an embarrassingly parallel problem. In practice, we divide sampled edge variables into batches, and calculate their conditional log-likelihoods independently on different cores. This leads to wall-clock time reductions, as the pseudo-likelihood calculation time scales with the inverse of the number of available cores. This (combined with edge variable sampling) can make the MPLE an attractive choice for very large valued networks, especially when many cores are available.

Taken together, the above computational techniques allow the MPLE to be used even for very large networks with highly dispersed counts (although not all of them are needed when counts are less variable, or on smaller networks). As we show, valued MPLE is very fast, and the resulting estimator can have low bias and high accuracy for valued networks; it offers high-quality calibration of uncertainty when the edge variance is large, but is prone to overconfidence for dependence terms (i.e., underestimation of the standard error) and conservative for nondependence terms when the edge variance is small.

Table 1
Network descriptive statistics of the studied cases.

	Network size	Binary density	Edge value			
			min	max	mean	std. dev.
Large-variance small network	33	0.50	0	3862	46.15	200.54
Small-variance small network	33	0.50	0	8	1.64	1.95
Small-variance large network	100	0.41	0	9	1.32	1.84

4. Study design

We evaluate the above estimation techniques via a parameter recovery study, in which we generate networks from a realistic generative model based on an initial fit to real-world social networks, estimate models to the simulated draws using each respective technique, and then examine the properties of the resulting estimators. Our generative model was created by fitting a valued ERGM to an empirical case (see below) using MCMLE; we then obtained 500 high-quality draws from the fitted ERGM using MCMC. For each draw, we obtained point and standard error estimates from each of the three study methods (MCMLE, CD, MPLE), evaluating the results with respect to wall-clock estimation time, bias, variance of the estimator, overall accuracy, and calibration (accuracy of estimated standard errors and confidence coverage). All modeling and analysis was performed using `statnet` (Handcock et al., 2008), specifically using the `ergm` 4.2.1 (Hunter et al., 2008b; Krivitsky et al., 2022), `ergm.count` 4.1.1 (Krivitsky et al., 2012), and `sna` 2.6 (Butts, 2008) libraries. Our MPLE implementation also made use of the `Rcpp` library (Eddelbuettel et al., 2011). The following subsections detail the data and model used, the setup of the estimation procedures, and the performance metrics.

4.1. Case study and model definition

To examine the performance of estimation methods for data with different network sizes and edge value variances, we construct the following study cases. They are based on real-world datasets of migration flows between counties in two U.S. states (New Mexico and North Carolina) (U.S. Census Bureau, 2018). The New Mexico data consists of 33 nodes and the North Carolina data 100 (henceforth the "small" network and the "large" network, respectively). For the New Mexico data, we generate two networks with different edge value distributions. The large-variance case uses the count of migrants between each directed county pair as the edge value, which ranges in integers from 0 to 3862 with standard deviation 201. For comparison, the edge value of the small-variance case takes the natural logarithm of migrant count (plus one), rounded to the nearest integer. Its edge value ranges from 0 to 8, with standard deviation 2. Ideally, we would generate large-variance and small-variance cases for the large network as well. Unfortunately, the large-variance large-network case turned out to be prohibitively computationally expensive for a simulation study comparing all standard methods, and we hence just include the small-variance large network case; the edge value is generated by the same manner discussed above, and the distribution is similar to the small network case, ranging from 0 to 9 with standard deviation 2. Table 1 displays the descriptive statistics of the three study cases.

Our ground-truth model is created by fitting a count-valued ERGM to the above networks. The sufficient statistics include a Sum term (intercept, the summation of all edge values), a Nonzero term (the count of nonzero edges), three exogenous covariates, and two dependence terms. The exogenous covariates are the population sizes of the sending and receiving counties (called Nodeocov and Nodeicov respectively), and the distance between counties (Edgecov) (all on natural log scale). The first dependence term, mutual, measures the reciprocity of the network, defined as

$$g_{\mathrm{m}} = \sum_{(i,j)\in\mathbb{Y}} \min\{y_{ij}, y_{ji}\}$$

The second dependence term, flow, is adapted from a previous model of inter-county migration networks in the United States (Huang and Butts, 2022). It calculates the summation of the volumetric flow of each node, which is the minimum of total inflow and total outflow for a node. It is a count-valued version of two-paths or mixed-two-star terms for binary networks (Morris et al., 2008). Formally,

$$g_f = \sum_{i\in\mathbb{V}} min\{\sum_{j\in\mathbb{V}\setminus i} y_{ij}, \sum_{k\in\mathbb{V}\setminus i} y_{ki}\}$$

where $\mathbb{V}$ is the vertex set. This model encompasses a diversity of different sufficient statistics commonly used in valued ERGMs, including graph-level baseline statistics (Sum, Nonzero), covariate effects (Nodeicov, Nodeocov, Edgecov), and dependence terms at the dyadic level (Mutual) and the triadic level (Flow). Although our aim here is to produce a deliberately simple model for purposes of evaluation (as opposed to a substantively detailed model of migration), our choice of statistics was informed by prior work on migration, and previous empirical analyses on migration-flow networks in particular (Boyle et al., 2014; Huang and Butts, 2022; Windzio, 2018; Zipf, 1946).

4.2. Methods under evaluation

As described above, we estimate parameters from the simulated network draws using three procedures: Contrastive Divergence (CD), Maximum Pseudo-Likelihood Estimation (MPLE), and Monte Carlo Maximum Likelihood Estimation (MCMLE). Estimation for each method was performed as follows.

For CD, we use the default settings of the `ergm.count` package, performing 8 Metropolis–Hastings steps, raising one proposal in each step. We also tried using more steps and/or more proposals within each step for CD. As is shown in the Appendix A, its estimation bias and calibration usually do not improve systematically as we increase the tuning parameters, and when it does, it fails to match other methods in comparable time. Therefore, we keep the most time efficient setting in our comparison.

For MPLE, we implement the procedure described in Section 3.3. To examine the performance of MPLE with various sample sizes, we consider three subsampled MPLE in each study case: the fast, the mid, and the full, which corresponds to uniformly sampling 50%, 75% and 100% of the edge variables in random, respectively; this corresponds to 528, 792, 1056 dyads for small networks, and 4950, 7425, 9900 dyads for the large network. We also consider the impact of multiple cores on execution time, calculating the wall-clock time for models estimated using 1,4, and 20 cores, respectively. In terms of edge support truncation, for small-variance cases, we use a uniform non-edgewise truncation; the support covers integers from 0 to λ times the max edge value of the network, where $\lambda = 4$ for the small network and $\lambda = 1.5$ for the large network as the latter has more information with more edge variables. Edgewise support truncation becomes a powerful tool to reduce computation time for large-variance networks; we set $\lambda = 4$ for the doubly-truncated edgewise support (see the second paragraph of Section 3.3.2). We also coerce the support to include integers from 0-value to the 80% quantile of the edge distribution for every edge variable, whose upper bound ranges from 43 to 48 and is typically close to the mean of the edge distribution. This is a conservative scheme for an edgewise support truncation with wide intervals and without coarsening, but is fast enough to have one order of magnitude less wall-clock time than MCMLE.

For MCMLE, we use the stochastic approximation method in `ergm.count`, a workhorse method that is also implemented and served as the default method in PNET (Wang et al., 2009). We made a

Table 2
Number of rounds for CD-MCMLE before convergence.

N of rounds	1	2	3	4	5	6	7
Count	434	38	17	9	0	1	1
Percentage (%)	86.8	7.6	3.4	1.8	0	0.2	0.2

few adjustments from the default setting to improve its performance based on the data structure of the study case and our exploratory experiments. First, we set the proposal distribution of Metropolis–Hastings algorithm in MCMC to be random, where every dyad has equal chance to be toggled. By default, the proposal distribution in `statnet` favors toggling nonzero edges than nulls, with the rationale that social networks tend to be sparse (for binary networks); we revoke this penalty towards nulls since the binary density of the network is high (0.4–0.5) in the study cases (removing the need for biased proposals), and the random proposal reduces the computational time of the MCMLE. Second, by default, the MCMC thinning interval is 1024 and the sample size of network statistics in each distribution returned by the algorithm is also 1024.[2] This setting is sufficient for the small-variance small network case, but for the other two cases, we increase these two parameters to be ten thousand; this helps with convergence and further increasing those parameters no longer brings performance gain based on our experiments. We evaluate the use of both CD and MPLE as seeding methods for MCMLE, referring to them as CD-MCMLE and MPLE-MCMLE, respectively. We employ default settings for CD, given that longer chains do not consistently enhance performance (see Appendix A). For the small-variance small network case, we use the "mid" MPLE for MCMLE seeding, but for the large-variance and the large network cases, we use the "fast" MPLE since the fast MPLE yields good-enough point estimations. It turns out that CD-MCMLE fails to converge for some of the large-variance cases; we detect this by examining the MCMC diagnostics plot, and reruning those cases until convergence (details in Table 2 in Section 5). We also turn off the bridge sampler that calculates the log likelihood to reduce the MCMLE computational time, since this is not involved in estimation.

All models were fit on a 44-core server, with 256 GB RAM. The processors are dual Intel Xeon E5-2699 2.2 GHz CPUs (22 cores/CPU). Estimation using R 4.1.1 was performed on Ubuntu 20.04.1. All procedures reported are based on a single core, except for the multi-core MPLE conditions.

4.3. Evaluation criteria

Since the methods of interest involve speed/quality trade-offs, it is necessary to evaluate these two dimensions simultaneously. To evaluate computational cost, we compute the wall-clock time for each method, as mean seconds needed to fit the target model to a simulated network. Since the speed of MPLE is dependent upon the number of parallel processes, we repeat the process using 1, 4, and 20 cores, respectively.

To evaluate estimation quality, we consider the bias, variance, overall accuracy, and calibration of each estimator using the following metrics.

We first compute the absolute relative bias (ARB) of estimators for each coefficient, using

$$ARB = \left| \frac{1}{m} \sum_{i=1}^{m} \frac{\hat{\theta}_i - \theta}{\theta} \right|$$

taking the average across m experiments, where $\hat{\theta}$ is the estimator and θ is the true value from the model that simulated the networks. The smaller the ARB, the less bias is introduced by the estimation procedure.

[2] We follow the default of `statnet` that sets MCMC burnin as 16 times the length of the thinning interval.

We also compute the variability for each estimator, via the (true) standard error of the estimated coefficient. Formally,

$$SE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (\hat{\theta}_i - \frac{1}{m} \sum_{j=1}^{m} \hat{\theta}_j)^2}$$

The smaller the variability, the more efficient and more precise the estimator is.

While bias and variance are each important, we are also interested in the total accuracy of the estimator (the extent to which it deviates, on average, from the true value). We measure this via the root-mean-square error (RMSE) i.e.

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (\hat{\theta}_i - \theta)^2}$$

The smaller the RMSE, the more accurate the estimator is on average.

Finally, we consider how well calibrated each estimator is, in terms of the associated estimates of uncertainty. To evaluate the bias in our second moment estimate, we compare the real standard error se and the estimated standard error $\hat{se}$ using

$$Calibration = \log \left[\frac{1}{m} \sum_{i=1}^{m} \frac{\hat{se}_i}{se} \right]$$

A positive number suggests the method is conservative, while a negative number suggests the method is overconfident. We also examine confidence coverage, specifically the proportion of cases in which the nominal 95% confidence interval (CI) for each parameter actually covers the true coefficient. Specifically, the coverage rate is computed by

$$Coverage = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{\theta \in [\hat{\theta}_i - z \cdot \hat{se}_i, \hat{\theta}_i + z \cdot \hat{se}_i]\}$$

where $z = 1.96$ for 95% CI. The closer to 95% the coverage is, the better calibrated the estimate is. Coverage rates above 95% suggest that the method is conservative, while coverage rates less than 95% suggest that the method is overconfident.

5. Results

5.1. The small-variance, small network case

Starting from the simplest case of the small-variance small network, we display the performance of each method in Figs. 1 and 2. Panel A in Fig. 1 shows the absolute relative bias (ARB) of the coefficient estimates. It shows that all methods produce very small numerical biases, 3% or less across all covariates and methods. CD and fast MPLE introduces larger biases; but as the sample size of MPLE increases, its bias reduces and gets close to that of MCMLE, seeded by either CD or MPLE. This finding is consistent with research on binary ERGMs finding that the MPLE introduces little bias in parameter estimation (van Duijn et al., 2009; Schmid and Desmarais, 2017). The lack of appreciable bias is an encouraging sign, suggesting that point estimation of valued ERGMs for small-variance small network is easy to acquire using whichever method we tested.

We also evaluate the (im)precision or variability of estimation (sometimes called efficiency), i.e. the true standard error of each estimator. Panel B in Fig. 1 shows that the variability of the MPLE decreases with more edges sampled, and that full MPLE and MCMLE are the most efficient methods. In general, variations of estimators using all methods are close to each other, suggesting that they have similar efficiency.

We then evaluate the total accuracy of estimators using the root-mean-square error (RMSE). A more holistic metric, the accuracy measurement combines bias and variation of estimation evaluated above, and smaller RMSE is preferred. Panel C of Fig. 1 displays RMSE scores,

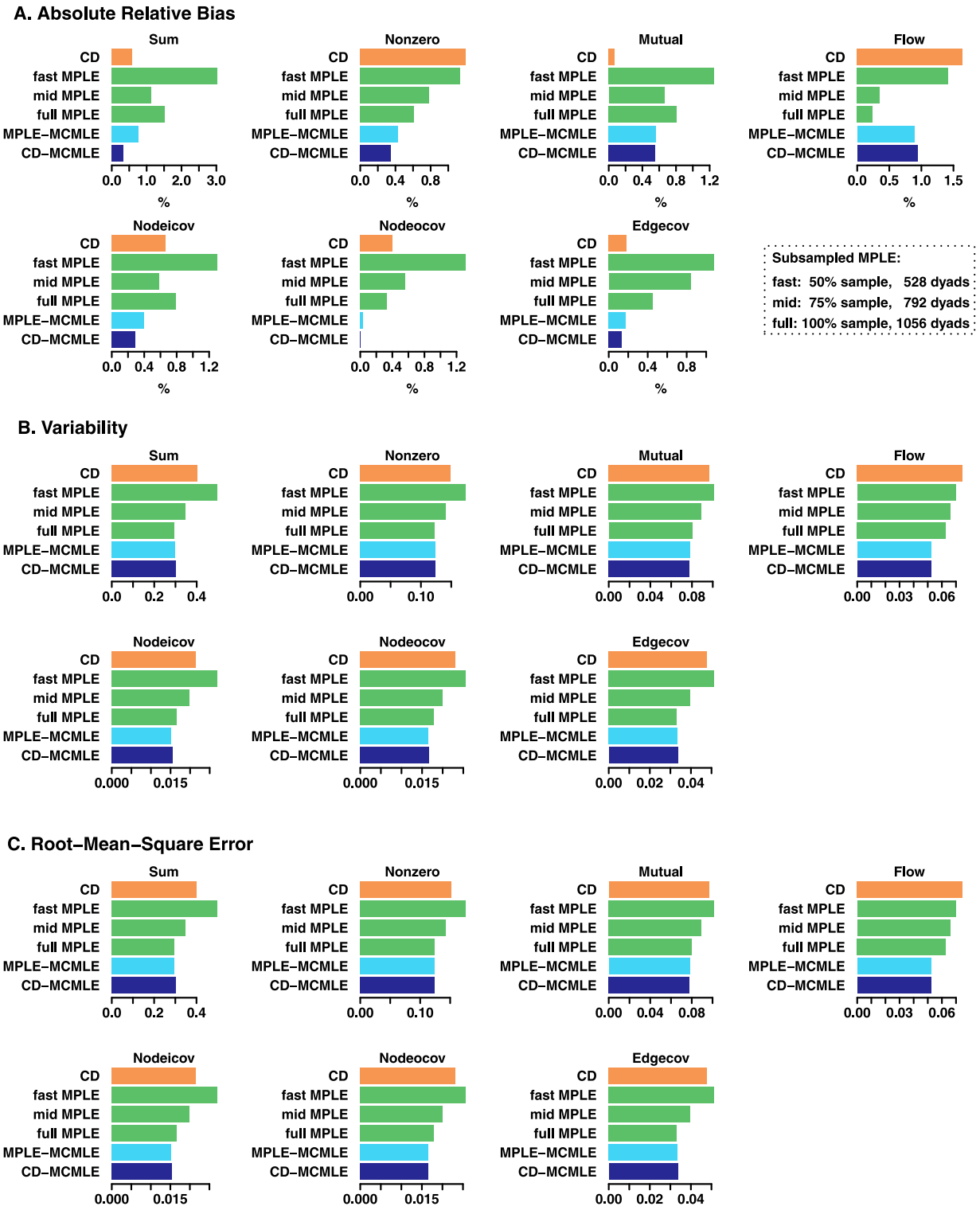


Fig. 1. Bias, variability, and RMSE of small-variance small network.

whose distribution is almost identical to the variability score in Panel B. The similarity between RMSE and variability reveals that biases contribute very little to the total RMSE, with accuracy being dominated by the performance in variability. Methods with good variability scores thus also have decent accuracy. Full MPLE and MCMLE has the smallest RMSE, though RMSEs for all methods under evaluation show only mild differences.

Besides performance in coefficient estimation, performance in estimating uncertainties is also evaluated, shown in the first two panels in Fig. 2. Panel D displays calibration of each method. An indicator of the bias in standard error estimation, a positive calibration score suggests overestimation of the uncertainty, and a negative calibration suggests underestimation. Noticeably, CD overestimates standard errors for all covariates by a large margin; CD's calibration scores are all above 1.9, indicating that the estimated standard error is more than 6.6 (i.e., $e^{1.9}$) times its true value. We experimented with different tuning parameters for CD, but could not find settings with both improved calibration and reasonable execution time (see Table A.2 in Appendix A). In

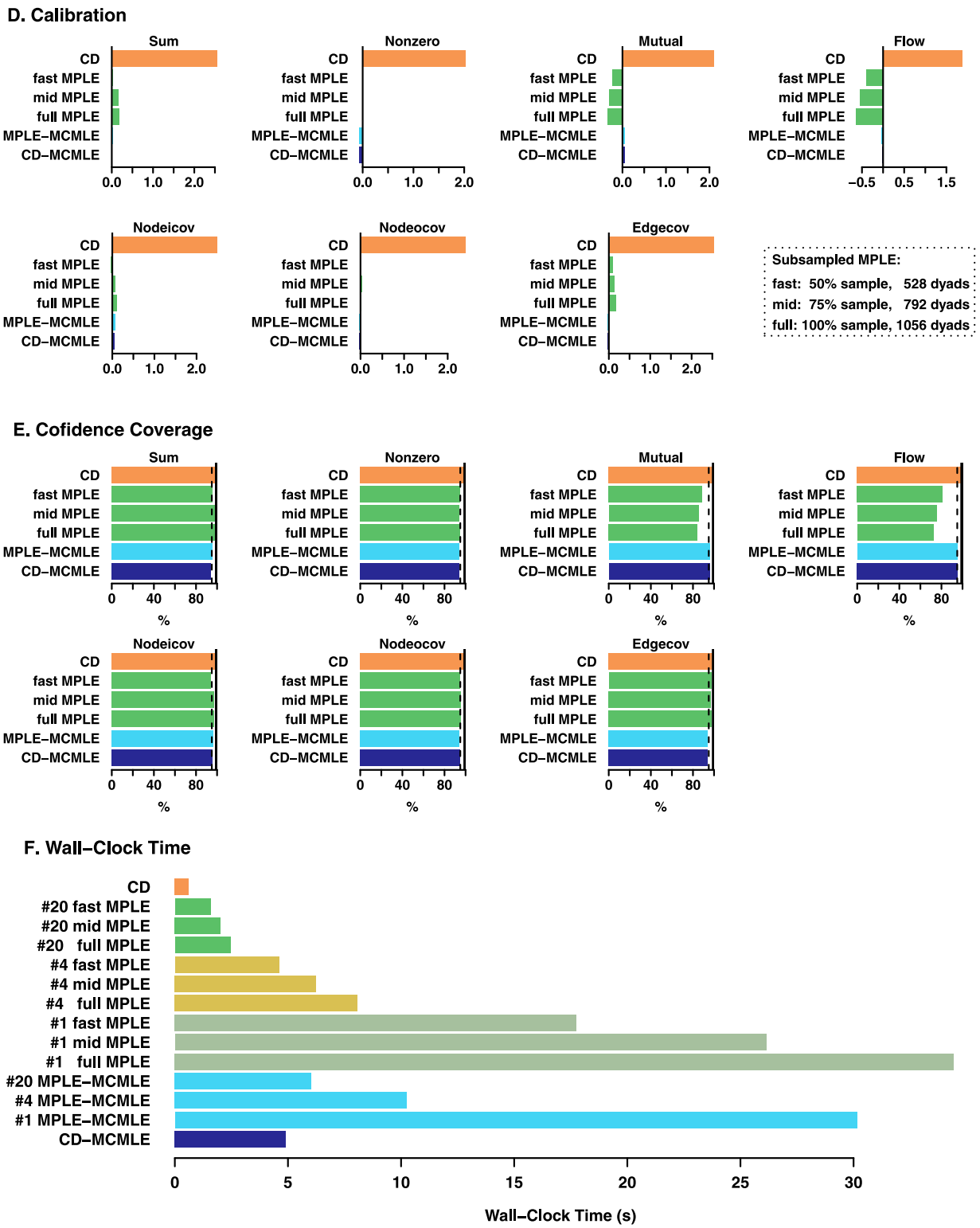


Fig. 2. Calibration, confidence coverage, and wall-clock time of small-variance small network.

summary, CD is too conservative to offer useful uncertainty estimations of covariates for small-variance small networks.

For MPLE, we find that it underestimates the uncertainties for the dependence terms (mutual and flow), but overestimates uncertainties for non-dependence terms, though the degree of inflation is very small. Previous simulation studies found similar patterns for MPLE on binary ERGMs (Lubbers and Snijders, 2007; van Duijn et al., 2009). Interestingly, the bias in standard error estimation increases with the sample size for MPLE. This is in part because the bias of estimation in statistical uncertainty is trivial when the numerical uncertainty is the main source of uncertainty for the MPLE (as is the case with small sample sizes); as the numerical uncertainties decrease with more edges sampled, the bias in statistical uncertainty becomes non-negligible. The calibration of MPLE is not as far off as that of CD, but its underestimation of standard errors is noticeable for dependence terms in the small-variance small network case.

While CD and the MPLE show varying degrees of error in standard deviation, both CD-MCMLE and MPLE-MCMLE have almost perfect calibration, suggesting that MCMLE is the best method for standard error estimation of small-variance small network.

Another metric that considers uncertainty estimation, confidence coverage is the proportion of model fits in which the true value of a given coefficient is covered by the estimated 95% confidence interval (CI), as is shown in Panel E of Fig. 2, where the dotted line is the 95% reference line and the solid line represents 100%. The figure tells a similar story to Panel D's calibration score, because confidence coverage performance is largely determined by performance in calibration of uncertainty when the bias of coefficient estimations is small. The figures show that CD overestimates standard errors so much that its CIs always cover the true value, making them conservative but uninformative. The MPLE's CIs cover the true values more than 95% for non-dependence terms, but under-cover the true values for the dependence term, with this deviation becoming larger as sample size increases. On the other hand, both CD-MCMLE and MPLE-MCMLE have coverage rates that are extremely close to 95%, showing its characteristic calibration advantage for small-variance small network of valued ERGMs estimation.

Lastly, Panel F of Fig. 2 displays the wall-clock time of each method. As expected, the wall-clock time for computing the MPLE can be greatly compressed by using a sample of edges to approximate the joint pseudo-likelihood function, or by using multiple cores to calculate conditional likelihoods. The fastest methods are CD followed by MPLE using 20 processors. Overall, the wall-clock time for different methods is short and varies modestly for this simple computation case, costing half a minute at most. Interestingly, while MCMLE is commonly believed to be a slow method, it is very fast in this simple case.

In summary, for small-variance small network data, all methods offers accurate and minimally-biased point estimates. CD offers uninformatively conservative uncertainty estimates, and MPLE's uncertainty estimation for dependence terms is noticeably overconfident. All methods are reasonably fast in this regime. With great performance in all metrics, MCMLE is an ideal method for valued ERGMs estimation for small-variance small network data.

5.2. The small-variance, large network case

The small-variance large network case has similar edge value distribution to the previous small network case, but its network size is 3 times bigger, meaning that its dyad count is 9.4 times the count of the previous case. Comparing these two cases offers insights about the influence of network size on estimation performance for each method.

Fig. 3 shows that, again, all methods introduce very little bias. One noticeable difference is that the bias of subsampled MPLE gets smaller in comparison with other methods. This suggests that the absolute number of edge variables sampled influences the performance of subsampled MPLE; for small networks, one needs to sample a larger proportion of edges, while for large networks, the percentage can be lower. This means that for the large network case, the fastest MPLE is already less biased than CD (whose bias from a larger tuning parameter setting gets outperformed by following up CD with MCMLE, see Table A.1 in Appendix A); this translates to the less biased performance of MPLE-seeded MCMLE, compared to the CD-seeded MCMLE as the figure shows. Panels B and C in Fig. 3 reveal almost identical patterns in variability and RMSE compared to the small-variance small network case. Variability decreases as MPLE's sample size increases, and all methods share similar variability. The RMSE distributions resembles those of the variability as the bias of all methods are largely ignoreable.

Comparing Panels D and E in Fig. 4 with those in the previous Fig. 2 suggests that patterns of uncertainty estimation performance across methods are generally invariant to network size. We again find that CD greatly overestimates uncertainty, leading to over-coverage of the confidence intervals; MPLE underestimates the uncertainty for dependence terms and the confidence intervals under-covers their true values. MCMLE offers great uncertainty estimation again, although, on close inspection, CD-MCMLE shows a (very) small tendency towards overconfidence that the MPLE-MCMLE lacks.

Panel F in Fig. 4 shows that the wall-clock time of MPLE and MCMLE scales with the network size, while that does not apply to CD. Subsampling and use of multiple cores effectively reduce the computational time of MPLE; this advantage becomes larger with network size, since larger networks require a greater share of computing time to be used for change score calculation. Since changescore calculations are embarrassingly parallelizable, gains from multi-core calculations grow accordingly in this regime. Overall, this scenario shows clear superiority of CD for computational time, followed by MPLE using subsampling and multi-core strategies. MCMCLE becomes quite slow here (with mean times between $\approx$13 min and roughly half an hour), making speed a potentially important consideration.

To recap, comparing the two small-variance cases with small and large network sizes suggests that the estimation quality of each method is largely invariant to the network size. The larger number of edge variables for the large network means that subsampled MPLE requires a smaller share of edge variables for good performance, and fast MPLE becomes a less biased and a better seeding method than CD for MCMLE (though the difference is small in the study cases). All methods have good first-order performance, though MCMLE is clearly superior for calibration (with CD being unacceptably poor). The major performance difference coming from the network size is that the superiority in computational efficiency for CD and multi-core subsampled MPLE becomes substantial when the network size grows.

5.3. The large-variance, small network case

Given the consistency seen in the two small-variance cases, we might expect the large variance, small network case to behave similarly. However, we observe very different results in the small network case when the variance of edge values becomes large. First of all, we observe that CD-MCMLE simply fails to converge for some of the simulated networks, as reflected by their MCMC diagnostics plots; here we follow the common procedure of rerunning them until convergence, though failure to attend to diagnostics could lead to problems in casual use. Table 2 summarizes the number of rounds CD-MCMLE went through before seeing convergence. Overall, it took 1.22 rounds on average for CD-MCMLE to converge (with a few cases taking more than five). The following results are based on their final (converged) rounds, as estimators from the failed rounds were very far from the true values.

Panel A in Fig. 5 shows that CD introduces relatively larger biases for the large-variance case, 5.2% and 3.8% for the nonzero and the flow terms, respectively. Although those biases are arguably not huge, they lead to failures in convergence for CD-MCMLE. On the contrary, MPLE introduces very little bias in its estimates, the largest bias of 1.8% coming from the fast MPLE for the flow term. That makes it an excellent seeding method, and indeed all MPLE-MCMLE models converged in their first attempt. This signifies that, similar to the binary ERGM scenario, MCMLE for valued ERGMs is sensitive to the seeding quality, and small improvements in biases of the seeding methods can make a difference. Another notable feature is that MPLE actually outperforms MCMLE in the bias metric, with the latter having a bias of about 5% for the triadic dependence term (flow).

Panel B in Fig. 5 shows that CD generally has larger variability than other methods, especially for the nonzero term. MPLE's variability decreases with more edge variables sampled, and gets close to MCMLE when all edges are utilized. Panel C in Fig. 5 demonstrates that, again, when biases are generally small, the accuracy metric resembles that of the variability.

Panel D in Fig. 6 shows that CD substantially overestimates the standard errors, so overly conservative uncertainty estimation is a consistent behavior for CD across all network size and edge variance

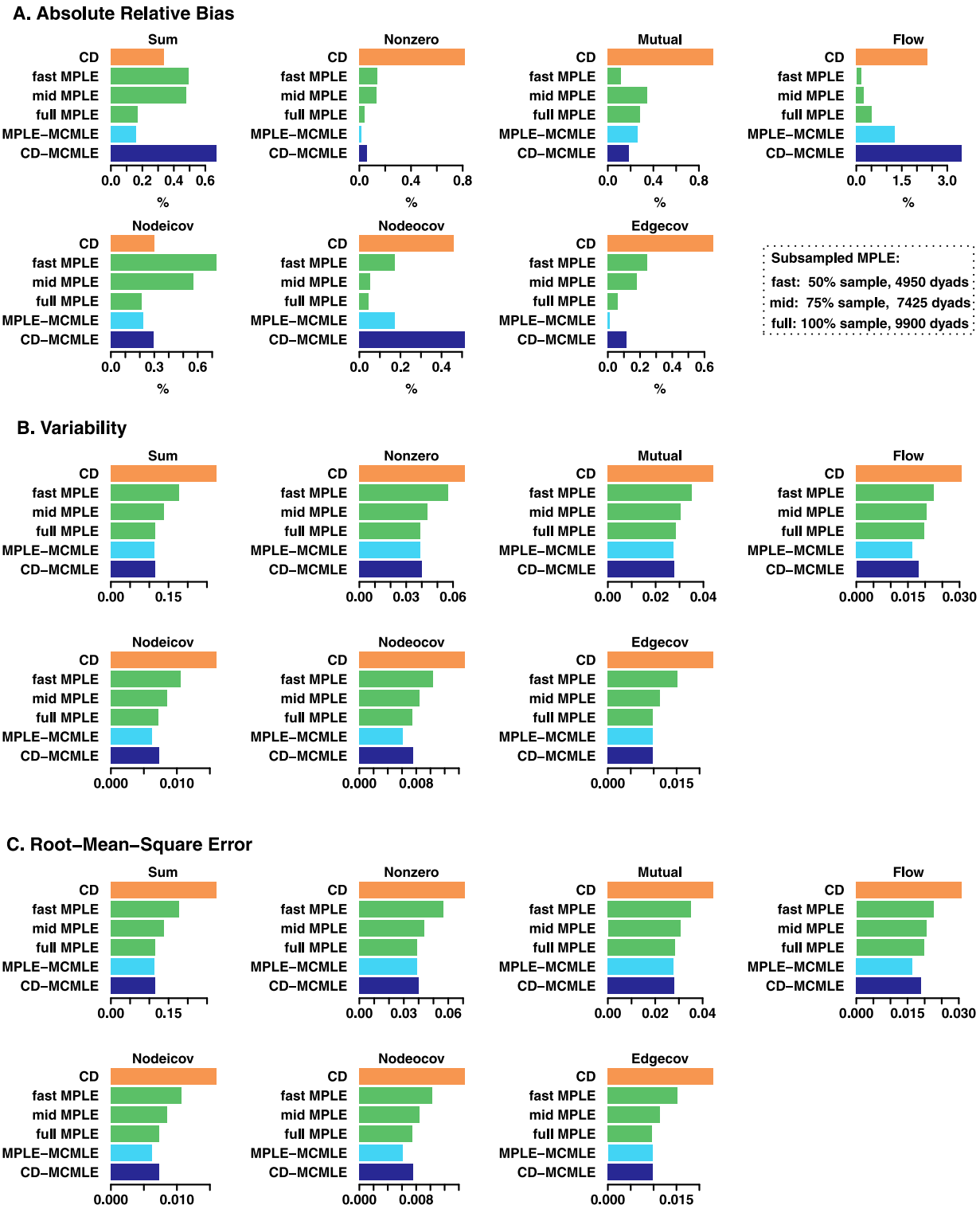


Fig. 3. Bias, variability, and RMSE for small-variance large network.

structures studied. By contrast, MPLE no longer suffers from calibration difficulties for the large-variance case, producing high-quality standard error estimates that match the performance of MCMLE, seeded by either CD or MPLE. Note that for the nonzero term, CD did not return an estimate for 22 of the 500 simulated networks, meaning that the calibration of CD for the nonzero term could actually be worse; we did not rerun them, as CD's uncertainty estimates are not useful even when the algorithm did converge. Panel E in Fig. 6 reveals that CD's confidence intervals have the over-coverage issue for every covariate except the nonzero term, which suffers from a larger first-order bias. MPLE and MCMLE, both with small bias in point estimation and uncertainty estimation, offer confidence coverage very close to the 95% benchmark.

Lastly, Panel F in Fig. 6 shows that for the large-variance network, MCMLE is an order of magnitude slower than CD and MPLE. Subsampling and use of multiple cores are still effective ways of reducing computational time for MPLE, and MPLE with 20 processors becomes even faster than CD. We should also note that the time reported here

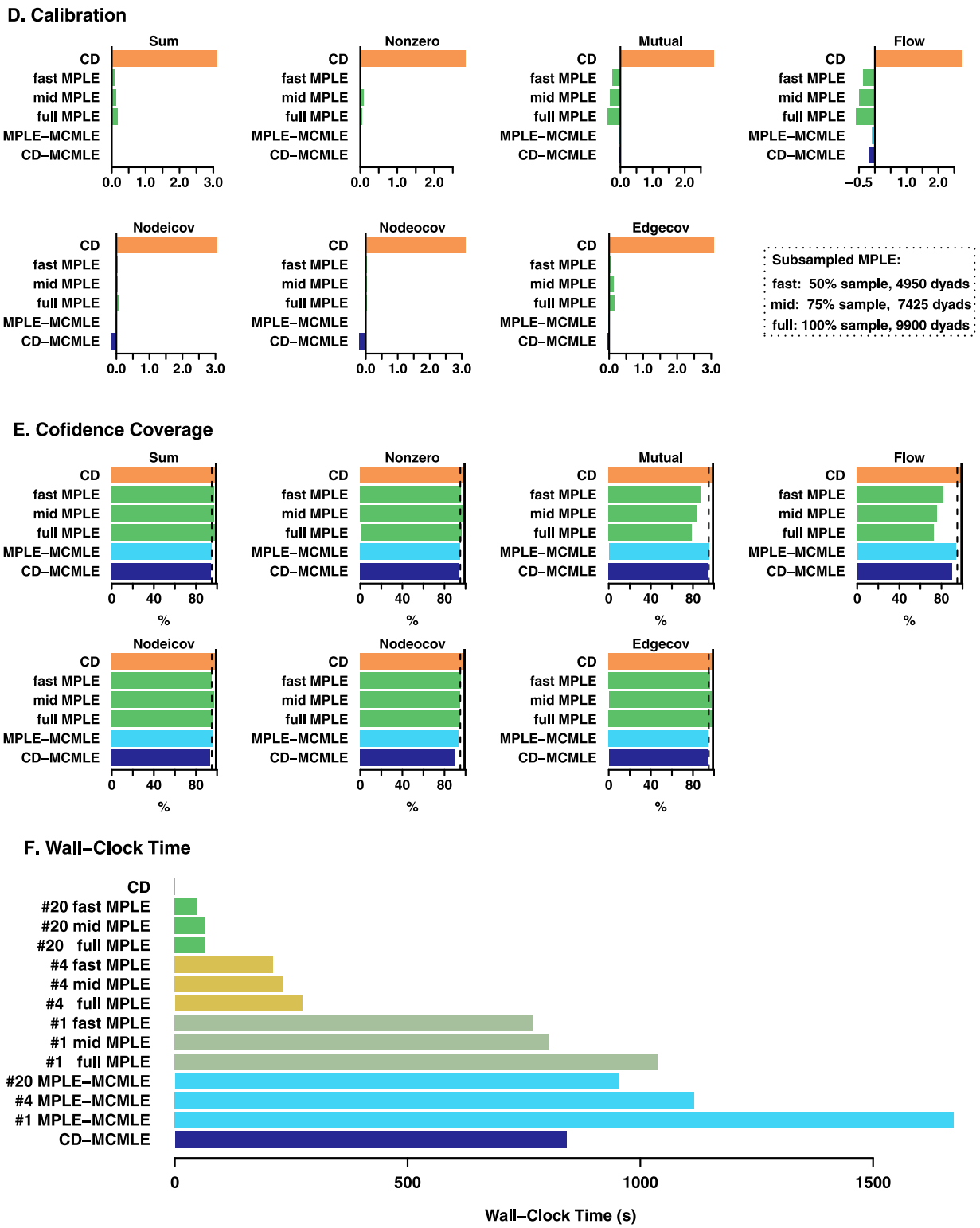


Fig. 4. Calibration, confidence coverage, and wall-clock time for small-variance large network.

for CD-MCMLE is the simple summation of their all rounds of wall-clock time. In reality, users need to spend more time digging into the diagnostics of MCMC after each round, and this makes CD-MCMLE even more slower than running MPLE-MCMLE.

Overall, we observe that the variance of edge values makes a substantial difference in the behavior of these estimation methods. CD generates fair point estimates in a speedy manner (albeit less accurate than its peers), but its calibration is overly conservative to the point of being unusable. The larger biases of CD estimators sometimes prevents convergence of CD-MCMLE, making MPLE a better seeding method in this scenario. MCMLE, seeded by either CD or MPLE, once converged, offers high-quality estimates at correspondingly high cost. Strikingly, however, we find that in this case MPLE generates estimators that match MCMLE in all metrics and introduces even less bias. Considering that MPLE is also an order of magnitude faster than MCMLE, it is clearly the superior method for the large-variance case. If higher-

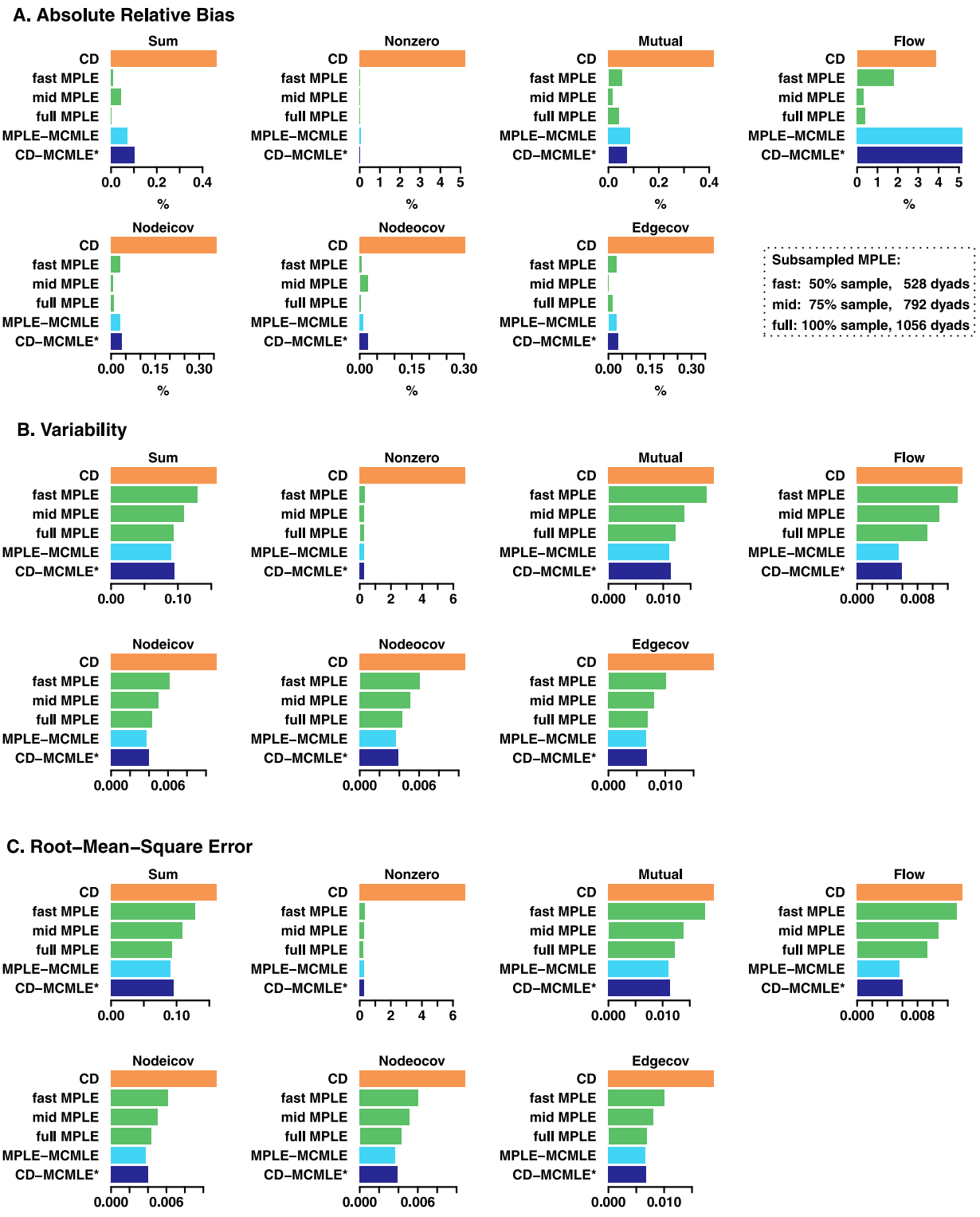


Fig. 5. Bias, variability, and RMSE of large-variance small network. *Note:* CD-MCMLE* results are from their final rounds with convergence.

quality estimates are needed, one can increase the sample size of MPLE or follow it up with MCMLE, though the former is a much faster option.

6. Discussion

Overall, this comparative simulation study reveals that the variance of the edge value makes a substantial difference in the performance of estimation methods for valued ERGMs, while network size primarily impacts computational cost. For small-variance data, all methods perform very well in point estimation, while CD greatly overestimates the uncertainty, and MPLE underestimates uncertainties of dependence terms. MCMLE seeded by either CD or MPLE offers high-quality estimates under all metrics. Wall-clock time of the methods are close to each other for small-variance small networks, but the speed advantages over MCMLE gets larger for CD and subsampled/multi-core MPLE as the network size increases. For very large graphs (especially when large numbers of cores become available), the relative speed advantage of the

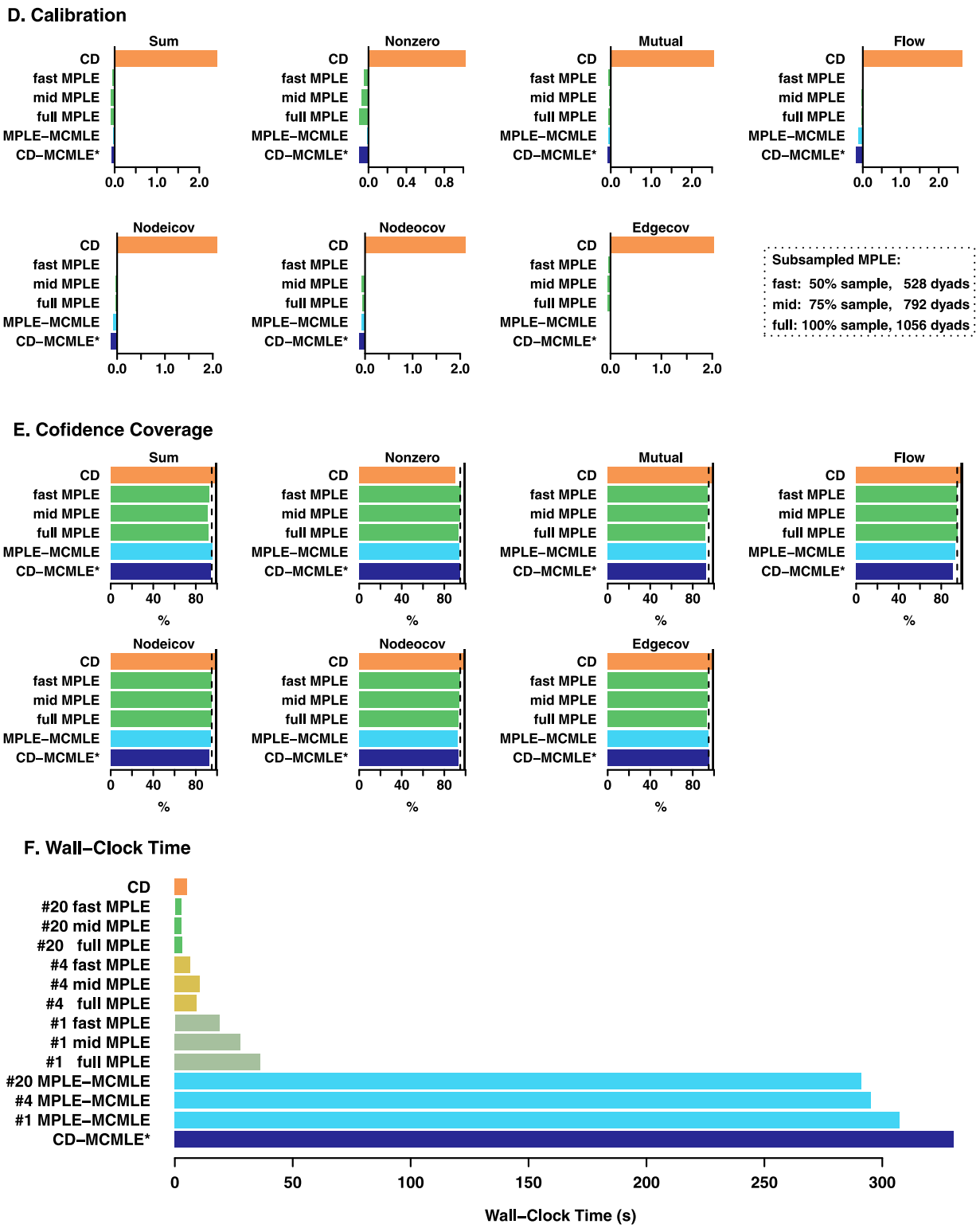


Fig. 6. Calibration, confidence coverage, and wall-clock time of large-variance small network. *Note:* CD-MCMLE* results are from their final rounds with convergence. Its wall-clock time is simple summation of all rounds of computation.

MPLE can become substantial, and may be a reasonable consideration in method selection.

For large-variance data, CD fails to offer reliable uncertainty estimates; its relatively high bias compared to MPLE also makes CD-MCMLE more prone to convergence failure. Both MPLE and MCMLE (given convergence of the latter) perform well in all quality evaluations, but MPLE is an order of magnitude faster than MCMLE, and can be further sped up by subsampling and parallel computing. Again, we observe that the speed advantages of MPLE become larger as graph size and edge variance increase.

The results suggest that for small networks with low-variance edges, MCMLE continues to serve as the estimation method of choice: it delivers high-quality point estimates and excellent calibration while still being computationally accessible in this regime. As network size

increases, MCMLE becomes increasingly cumbersome, as its computational time scales up faster than other methods. When MCMLE is too slow to perform well, CD and MPLE can serve as useful tools for tasks that only require first-order point estimation, such as exploratory analysis, prediction, or generating models for network simulation. The subsampling and multi-core features of MPLE offer useful tools for fast computation of models on large networks, and the numbers of edge variables that must be sampled for strong performance are an increasingly small fraction of the total dyad count as network size grows, further enhancing its computational advantages. Although MPLE calibration is certainly good enough to be useful (especially for independence terms), we clearly see the tendency towards overconfidence in dependence terms found in binary ERGM studies, and nominal confidence intervals for these terms are likely to be too small. Analyses relying on coverage for such terms should be regarded as heuristic (though the maximum extent of miscalibration seen here may provide some guidance with respect to the degree of error that could be present).

For networks with large edge variance, MPLE could be the go-to method, yielding accurate estimators with good calibration. Besides subsampling and parallel computing choices, our implementation of edgewise support truncation further offers MPLE an edge in speed without compromising its estimation quality (as edgewise and nonedgewise truncation of MPLE offers commensurate results in quality metrics for large variance data). This enables the use of MPLE to estimate valued ERGMs that were previously blocked by computational barriers, such as high-dimensional models on large networks with high edge variance. It also offers a flexible framework to make trade-offs between estimation quality and computational time, by tuning the sample size of the edge variables and the structure of the edge support truncation. While MPLE has been seen as a sub-optimal choice for binary ERGMs (van Duijn et al., 2009), this comparative simulation study reveals an area where it can be the effective approach.

Our experiments also offer insights about MCMC-based estimation methods. This paper shows that MCMLE for valued ERGMs depends on high-quality starting values, especially for large-variance networks. Both CD and MPLE are useful tools for MCMLE seeding, but the relatively larger biases of CD for large-variance data makes CD-MCMLE more fragile. One potential reason for this fragility is the difficulty of CD in reproducing the dichotomized density of the networks (i.e. the proportion of nonzero edges). With larger ranges of edge values, the toggling of values between zero and one become more unlikely, and the difficulty of matching the target density increases. (This is, of course, a special case of zero-inflation, a common phenomenon in count data models beyond network settings.) This is reflected by the phenomenon that the sufficient statistic consistently observed to fail in CD-MCMLE was the nonzero edge count in the large-variance experiment. The issue can be worsened by the mismatch of the MCMC algorithm design and typical properties of valued networks. Compared to their binary counterparts, valued social networks are frequently denser (in the dichotomized density). However, MCMC algorithms in existing software are optimized for sparse, unvalued social networks. We observed improvement in computational time when switching from an MCMC proposal that favors toggling nonzero edges (so-called TNT, or more accurately "tie-random dyad") to a random proposal, which is the one that offers the highest likelihood of toggling empty edges among existing algorithms. To improve the performance of MCMC-based methods for valued ERGMs, future research could consider experimenting with MCMC algorithms that pay more attention to the toggling between value zero and one, such as proposals that favors toggling zero-value edges.

As with any simulation study, one trades off the "realism" of performance on a realistic case against some degree of generality. Although we vary the network size and edge value variance to emulate different application settings, we cannot rule out the possibility that some methods studied here may perform better or worse under other conditions. We encourage future research using simulation studies based on different use cases and model specifications, including non-Poissonian reference measures. Given that we find that the edge value variance plays an important role in influencing the performance of valued ERGM estimation, it would be of interest to experiment with more fine-grained classification of the scale of edge variance, in search of an empirical rule of thumb for when MCMLE or MPLE would be the better choice.

Our study also suggests the continuing relevance of the MPLE to ERGM methodology. Our implementation of MPLE for valued ERGMs enables estimation for large-variance data in feasible time and with high-quality results. With good overall accuracy, high speed, and flexible tunability, MPLE would be an excellent general use estimation method for valued ERGMs if its calibration could be improved for small-variance networks. Our findings suggest the value of work on methods that may help further improve calibration of MPLE in the count-valued case; such advances may build on methods that shown to help calibration for binary ERGMs, such as bootstrap resampling (Desmarais and Cranmer, 2012b; Schmid and Desmarais, 2017) and regularization (van Duijn et al., 2009).

Lastly, we should emphasize that it is important to perform model evaluation after estimation for generative network models like ERGMs. While this should be a standard procedure regardless of the estimation method, it is an especially important reminder as the field observes the revival of non-simulation and local-simulation methods such as MPLE and CD, thanks to emerging methodological innovations and new data structures. These methods are less prone to convergence failures, which can have the hidden liability of making it harder to spot poorly-behaved models (an issue encountered in the early use of the MPLE before the availability of simulation-based evaluation, as discussed by Snijders (2002)). We recommend that researchers evaluate model adequacy by simulating networks from the fitted model and comparing their key network features and specified sufficient statistics with their observed counterparts, e.g. following the procedure of Hunter et al. (2008a), where feasible. Fortunately, simulation-based evaluation is computationally much cheaper than simulation-based estimation, as the former only requires simulation from one model while the latter needs to explore a set of models in the parameter space; thus, even when MCMLE is computationally prohibitive for evaluation, MCMC adequacy checks (a.k.a. goodness-of-fit checks) are often feasible. For sufficiently large, high-variance systems in which even this is infeasible, alternative checks are needed. Although this regime remains an open problem, conditional simulation using e.g. Held-Out Predictive Evaluation (Wang et al., 2016; Yin et al., 2019) may be one useful approach, provided that enough dependence-graph adjacent edge variables are held out simultaneously to permit detection of degeneracy. Some work has been done on bounding techniques for dichotomous networks that can in some cases rule out degeneracy without resorting to simulation (Butts, 2011); it is unclear whether similar techniques can be developed for count-valued networks, but if so such methods could prove useful where simulation is impractical. In general, evaluation for networks that are too large for complete simulation (in the count-data case or otherwise) is an important frontier for future work.

7. Conclusion

ERGMs, especially for valued networks, can be computationally expensive to estimate. In searching for a fast and reliable computational method, we implemented MPLE for count-valued ERGMs, and performed a comparative simulation study using three methods: CD, MCMLE, and MPLE. We found that the edge value variance is critical in determining the performance of computational methods for valued ERGMs, while the network size mainly influences their relative merit in computational efficiency. For small-variance networks, point estimates are easy to acquire using whichever method, while CD greatly overestimates uncertainties and MPLE underestimates them for dependence

terms. All methods have similar wall-clock time. For large-variance networks, both MPLE and MCMLE offer strong performance for estimating both coefficient and uncertainties, although MPLE is an order of magnitude faster than MCMLE.

On the basis of this study, we recommend that researchers pay attention to the variance of edge values in choosing computational methods. For small-variance data, MCMLE should be the default method where feasible, although CD is useful for point estimations; MPLE is suited for large networks and high-dimensional models, especially with a large number of available processors, but caveats should be given for interpreting its standard errors for dependence terms. For large-variance networks, MPLE is a solid method, and researchers can design the size of edge sample and the structure of edge support truncation based on the computational resources at hand and the requirements of estimation quality. Our experiments also demonstrate that both CD and MPLE are useful tools for MCMLE seeding, although CD is better for simpler cases with its speed advantage and MPLE is better able to offer high-quality seeds across all scenarios.

In summary, with insights about the behaviors of each method under different network sizes and edge variances, this paper offers a guideline for choosing and tuning computational methods for valued ERGM estimation. The implementation of a flexible subsampled parallelizable MPLE framework is demonstrated to be a powerful tool; we envision it to empower researchers with large-variance big network data and high-dimensional model design, freeing them from the need to employ data-reduction and model-simplification compromises because of computational constraints.

Funding information

This work was supported by National Science Foundation, United States award SES-1826589 and NIH, United States award 1R01GM144964-01.

Acknowledgments

We thank Katherine Faust, David Schaefer, and attendees of the Social Networks Research Group at UC Irvine for their comments and suggestions.

Appendix A. Contrastive divergence with different parameters

Results in Section 5 showed that despite its time efficiency, CD has two limitations. First, its bias is larger than subsampled MPLE, making it a suboptimal seeding method for MCMLE (especially when the edge variance is large). Second, its calibration of uncertainty is overly conservative, making it an uninformative method for second moment estimates. This leads to the question of whether one could tweak its tuning parameters to trade its time efficiency for less biased and better calibrated estimations. In this regard, we study the quality of CD estimators when we vary CD's major parameters: steps and multiplicity. "Steps" determines the number of Metropolis–Hastings steps, and "multiplicity" determines the number of proposal for each step. The default setting in `ergm.count` package for CD is 8 steps and 1 multiplicity, which was the setting reported in Section 5. Here we compare that with different combinations of modified tuning parameters.

Table A.1 shows the bias and the wall-clock time for CD under different tuning parameters versus MPLE under its configuration for MCMLE seeding using a single processor. In this section, we only report performance for the dependence terms because of space limitation, but estimators for other covariates generally share similar patterns. For both small-variance small network and large-variance small networks, increasing either steps or multiplicity or both does not monotonically reduce the bias, while the wall-clock time increases monotonically as expected. For this reason, the default CD configuration seems to be the optimal choice as a seeding method for MCMLE. For the small-variance large network, although larger multiplicity does not bring less bias, increasing steps does brings monotonic decrease in bias estimation. However, CD's bias is always larger than that of MPLE, even when its wall-clock time surpasses MPLE's. This suggests that for small-variance large networks, one can increase steps to reduce the bias of CD, but it is not as efficient as using MPLE instead, which offers better estimators with less time used. To recap, increasing either or both tuning parameters for CD usually fails to yield less biased estimators, and even when that does, it is not as time-efficient as using MPLE.

Table A.1
Bias and time of CD vs. MPLE.

	Contrastive divergence						MPLE
Steps	8	80	8	80	800	8000	
Multiplicity	1	1	10	10	1	1	
Small-variance small network							
Bias(Mutual) (%)	0.07	0.64	1.07	1.31	0.64	0.53	0.66
Bias(Flow) (%)	1.64	1.38	0.51	0.68	2.54	3.53	0.35
Wall-clock time (s)	0.60	1.92	2.01	21.55	30.77	342.95	26.14
Small-variance large network							
Bias(Mutual) (%)	0.93	0.55	1.66	0.60	0.34	0.30	0.12
Bias(Flow) (%)	2.35	1.10	1.98	1.17	0.89	0.72	0.15
Wall-clock time (s)	1.15	6.00	6.34	69.82	96.63	2177.46	769.77
Large-variance small network							
Bias(Mutual) (%)	0.42	79.94	87.41	99.4	139.9	116.13	0.05
Bias(Flow) (%)	3.86	591.9	469.13	261.75	97.59	29.45	1.78
Wall-clock time (s)	5.22	12.57	13.02	94.71	81.46	876.71	18.83

Note: We use the chosen seeding setting for MPLE: 50% sample sizes for large-variance and large-network data, 75% for small-variance small network data. Wall-clock time of MPLE is from the slowest setting using one core.

Table A.2
Calibration and time of CD vs. MCMLE.

	Contrastive divergence						MCMLE
Steps	8	80	8	80	800	8000	
Multiplicity	1	1	10	10	1	1	
Small-variance small network							
Mutual	2.11	1.04	2.73	1.75	0.22	0.05	0.04
Flow	1.88	0.93	2.57	1.51	0.19	−0.02	−0.03
Wall-clock time (s)	0.60	1.92	2.01	21.55	30.77	342.95	30.16
Small-variance large network							
Mutual	2.91	1.99	3.21	2.75	0.95	0.16	−0.03
Flow	2.76	1.88	3.10	2.53	0.98	0.21	−0.09
Wall-clock time (s)	1.15	6.00	6.34	69.82	96.63	2177.46	1673.59
Large-variance small network							
Mutual	2.55	0.97	1.06	0.36	−0.33	−0.56	−0.05
Flow	2.63	0.33	1.23	0.05	−0.70	−0.96	−0.11
Wall-clock time (s)	5.22	12.57	13.02	94.71	81.46	876.71	307.35

Note: MCMLE is seeded by MPLE under the specified setting in Results using one core.

Table A.2 displays the calibration for the two dependence terms and wall-clock time of CD under various tuning parameters versus the benchmark MCMLE, seeded by MPLE with a single processor. For both the small-variance small network and the small-variance large network, increasing multiplicity does not lead to better calibrated standard errors, but increasing steps is associated with monotonic improvement in calibration. Nonetheless, CD fails to offer as well-calibrated estimates as MCMLE in comparable time spans. This suggests that, for small-variance data, compared to increasing steps for CD, it is a better choice to directly use MCMLE for a time-efficient and well-calibrated second-moment measurement. For large-variance small-network data, although increasing steps and/or multiplicity in general alleviate its overestimation of uncertainty, increasing steps beyond a certain point can lead to underestimation of uncertainty. Regardless, its calibration never beats MCMLE. Overall, our experiments suggest that increasing tuning parameters beyond the default setting in `statnet` does not always improve its calibration, and when it does, it is not as time-efficient as using MCMLE for calibration.

Appendix B. Supplementary material

Replication data and the R source code for the MPLE can be found at http://dx.doi.org/10.7910/DVN/BFVXZ6.

References

Aicher, Christopher, Jacobs, Abigail Z., Clauset, Aaron, 2014. Learning latent block structure in weighted networks. J. Complex Netw. 3, 221–248.

Aksoy, Ozan, Yıldırım, Sinan, A model of dynamic flows: Explaining Turkey's inter-provincial migration, Sociol. Methodol., forthcoming.

Altman, Douglas G, Royston, Patrick, 2006. The cost of dichotomising continuous variables. Bmj 332, 1080.

An, Weihua, 2016. Fitting ERGMs on big networks. Soc. Sci. Res. 59, 107–119.

Anderson, C., Wasserman, S., Crouch, B., 1999. A p* Primer: Logit models for social networks. Social Networks 21, 37–66.

Asuncion, A., Liu, Q., Ihler, P.A., Smyth, 2010. Particle filtered MCMC-MLE with connections to contrastive divergence. In: International Conference on Machine Learning. ICML.

Bernard, H Russell, Killworth, Peter, Sailer, Lee, 1979. Informant accuracy in social networks IV: A comparison of clique-level structure in behavioral and cognitive network data. Social Networks 2, 191–218.

Besag, Julian, 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Stat. Methodol. 36, 192–225.

Bhamidi, Shankar, Bresler, Guy, Sly, Allan, 2011. Mixing time of exponential random graphs. Ann. Appl. Probab. 2146–2170.

Block, Per, Stadtfeld, Christoph, Robins, Garry, 2022. A statistical model for the analysis of mobility tables as weighted networks with an application to faculty hiring networks. Social Networks 68, 264–278.

Boyle, Paul, Halfacree Keith, H., Vaughan, Robinson, Vaughan, Robinson, 2014. Exploring Contemporary Migration. Routledge, Abingdon, United Kingdom.

Butts, Carter T., 2008. Social Network Analysis with sna. J. Stat. Softw. 24, 1–51.

Butts, Carter T., 2011. Bernoulli graph bounds for general random graphs. Sociol. Methodol. 41, 299–345.

Butts, Carter T., 2019. A dynamic process interpretation of the sparse ERGM reference model. J. Math. Sociol. 43, 40–57.

Butts, Carter T., 2020. A dynamic process reference model for sparse networks with reciprocity. J. Math. Sociol..

Butts, Carter T., Petrescu-Prahova, Miruna, Remy Cross, B., 2007. Responder communication networks in the world trade center disaster: Implications for modeling of communication within emergency settings. J. Math. Sociol. 31, 121–147.

Cranmer, Skyler J., Desmarais, Bruce A., 2011. Inferential network analysis with exponential random graph models. Political Anal. 19, 66–86.

Dekker, David, Krackhardt, David, Snijders, Tom A.B., 2007. Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. Psychometrika 72, 563–581.

Desmarais, Bruce A., Cranmer, Skyler J., 2012a. Statistical inference for valued-edge networks: The generalized exponential random graph model. PLoS ONE 7, e30136.

Desmarais, Bruce A., Cranmer, Skyler J., 2012b. Statistical mechanics of networks: Estimation and uncertainty. Physica A 391, 1865–1876.

Drabek, Thomas E., Tamminga, Harriet L., Kilijanek, Thomas S., Adams, Christopher R., 1981. Managing multiorganizational emergency responses: Emergent search and rescue networks in natural disaster and remote area settings. In: Number Monograph 33 in Program on Technology, Environment, and Man. Institute of Behavioral Sciences, University of Colorado, Boulder, CO.

Eddelbuettel, Dirk, François, Romain, Allaire, J., Ushey, Kevin, Kou, Qiang, Russel, N., Chambers, John, Bates, D., 2011. Rcpp: Seamless R and C++ integration. J. Stat. Softw. 40, 1–18.

Faust, Katherine, 2011. Animal social networks. In: The SAGE HandBook of Social Network Analysis, Vol. 148. p. 166.

Fowler, James H., 2006. Connecting the Congress: A study of cosponsorship networks. Political Anal. 14, 456–487.

Geyer, Charles J., Thompson, Elizabeth A., 1992. Constrained monte carlo maximum likelihood for dependent data. J. R. Stat. Soc. Ser. B Stat. Methodol. 54, 657–683.

Granovetter, Mark S., 1973. The strength of weak ties. Am. J. Sociol. 78, 1360–1380.

Handcock, Mark S., 2003. Statistical models for social networks: Inference and degeneracy. In: Breiger, Ron, Carley, Kathleen M., Pattison, Philippa (Eds.), Dynamic Social Network Modeling and Analysis. National Academies Press, Washington, DC, pp. 229–240.

Handcock, Mark S., Hunter, David R., Butts, Carter T., Goodreau, Steven M., Morris, Martina, 2008. statnet: Software tools for the representation, visualization, analysis and simulation of network data. J. Stat. Softw. 24, 1548–7660.

Hinton, Geoffrey E., 2002. Training products of experts by minimizing contrastive divergence. Neural Comput. 14, 1771–1800.

Hoff, Peter D., Raftery, Adrian E., Handcock, Mark S., 2002. Latent space approaches to social network analysis. J. Amer. Statist. Assoc. 97, 1090–1098.

Huang, Peng, Butts, Carter T., 2022. Rooted america: Immobility and segregation of the inter-county migration networks. arXiv preprint arXiv:2205.02347.

Hummel, Ruth M., Hunter, David R., Handcock, Mark S., 2012. Improving simulation-based algorithms for fitting ERGMs. J. Comput. Graph. Statist. 21, 920–939.

Hunter, David R., Goodreau, Steven M., Handcock, Mark S., 2008a. Goodness of fit of social network models. J. Amer. Statist. Assoc. 103, 248–258.

Hunter, David R., Handcock, Mark S., Butts, Carter T., Goodreau, Steven M., Morris, Martina, 2008b. ergm: A package to fit, simulate and diagnose exponential-family models for networks. J. Stat. Softw. 24, nihpa54860.

Hunter, David R., Krivitsky, Pavel N., Schweinberger, Michael, 2012. Computational statistical methods for social network models. J. Comput. Graph. Statist. 21, 856–882.

Hyvärinen, Aapo, 2006. Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. Neural Comput. 18, 2283–2292.

Krackhardt, David, 1988. Predicting with networks: Nonparametric multiple regression analyses of dyadic data. Social Networks 10, 359–382.

Krivitsky, Pavel N., 2012. Exponential-family random graph models for valued networks. Electron. J. Stat. 6, 1100–1128.

Krivitsky, Pavel N., 2017. Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models. Comput. Statist. Data Anal. 107, 149–161.

Krivitsky, Pavel N., Butts, Carter T., 2017. Exponential-family random graph models for rank-order relational data. Sociol. Methodol. 47, 68–112.

Krivitsky, Pavel N., Handcock, Mark S., Hunter, David R., 2012. Package 'ergm.count'. J. Stat. 6, 1100–1128.

Krivitsky, Pavel N., Hunter, David R., Morris, Martina, Klumb, Chad, 2022. ergm 4: Computational improvements. arXiv preprint arXiv:2203.08198.

Leal, Diego F., 2021. Network inequalities and international migration in the Americas. Am. J. Sociol. 126, 1067–1126.

Lubbers, Miranda J., Snijders, Tom A.B., 2007. A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. Social Networks 29, 489–507.

McMillan, Cassie, 2022. Worth the weight: Conceptualizing and measuring strong versus weak tie homophily. Social Networks 68, 139–147.

Mele, Angelo, 2017. A structural model of dense network formation. Econometrica 85, 825–850.

Morris, Martina, Handcock, Mark S., Hunter, David R., 2008. Specification of exponential-family random graph models: Terms and computational aspects. J. Stat. Softw. 24, 1548–7660.

Nowicki, Krzysztof, Snijders, Tom A.B., 2001. Estimation and prediction for stochastic blockstructures. J. Amer. Statist. Assoc. 96, 1077–1087.

Robins, Garry L., Pattison, Philippa E., Wasserman, Stanley, 1999. Logit models and logistic regressions for social networks, III. Valued relations. Psychometrika 64, 371–394.

Schmid, Christian S., Desmarais, Bruce A., 2017. Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap. In: 2017 IEEE International Conference on Big Data (Big Data). pp. 116–121.

Simpson, Sean L., DuBois Bowman, F., Laurienti, Paul J., 2013. Analyzing complex functional brain networks: Fusing statistics and network science to understand the brain. Stat. Surv. 7, 1–36.

Snijders, Tom A.B., 2002. Markov chain monte carlo estimation of exponential random graph models. J. Soc. Struct. 3, 1–40.

Strauss, David, Ikeda, Michael, 1990. Pseudolikelihood estimation for social networks. J. Amer. Statist. Assoc. 85, 204–212.

Tan, Linda S.L., Friel, Nial, 2020. Bayesian variational inference for exponential random graph models. J. Comput. Graph. Statist. 29, 910–928.

Ulibarri, Nicola, Scott, Tyler A., 2017. Linking network structure to collaborative governance. J. Public Adm. Res. Theory 27, 163–181.

U.S. Census Bureau, 2018. County-to-county migration flows: 2011–2015 ACS. https://www.census.gov/data/tables/2015/demo/geographic-mobility/county-to-county-migration-2011-2015.html. Accessed: 1/15/2019.

van Duijn, Marijtje A.J., Gile, Krista J., Handcock, Mark S., 2009. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. Social Networks 31, 52–62.

Vega Yon, G., Slaughter, Andrew, la Haye, Kayla de, 2021. Exponential random graph models for little networks. Social Networks 64, 225–238.

Vu, Duy Q., Hunter, David R., Schweinberger, Michael, 2013. Model-based clustering of large networks. Ann. Appl. Stat. 7, 1010–1039.

Wainwright, Martin J., Jordan, Michael I., et al., 2008. Graphical models, exponential families, and variational inference. Found. Trends® Mach. Learn. 1, 1–305.

Wang, Cheng, Butts, Carter T., Hipp, John R., Jose, Rupa, Lakon, Cynthia M., 2016. Multiple imputation for missing edge data: A predictive evaluation method with application to add health. Social Networks 45, 89–98.

Wang, Peng, Robins, Garry, Pattison, Philippa, 2009. PNet: Program for the Simulation and Estimation of Exponential RandOm Graph (P*) Models. The University of Melbourne.

Ward, Michael D., Ahlquist, John S., Rozenas, Arturas, 2013. Gravity's Rainbow: A dynamic latent space model for the world trade network. Netw. Sci. 1, 95–118.

Windzio, Michael, 2018. The network of global migration 1990–2013. Social Networks 53, 20–29.
Windzio, Michael, Teney, Céline, Lenkewitz, Sven, 2019. A network analysis of intra-EU migration flows: how regulatory policies, economic inequalities and the network-topology shape the intra-EU migration space. J. Ethn. Migr. Stud. 1–19.
Yin, Fan, Phillips, Nolan E., Butts, Carter T., 2019. Selection of exponential-family random graph models via held-out predictive evaluation (HOPE). arXiv:1908.05873.
Zipf, George Kingsley, 1946. The P1 P2/D hypothesis: On the intercity movement of persons. Am. Sociol. Rev. 11, 677–686.

Marginal-preserving Imputation of Three-way Array Data in Nested Structures, with Application to Small Areal Units[1]

Loring J. Thomas[1], Peng Huang[1,2], Xiaoshuang Iris Luo[3], John R. Hipp[3], Carter T. Butts[1,2,4,5,i]

[1] University of California, Irvine, Department of Sociology
[2] University of California, Irvine, Department of Statistics
[3] University of California, Irvine, Department of Criminology, Law, and Society
[4] University of Califorina, Irvine, Department of Computer Science
[5] University of Califorina, Irvine, Department of Electrical Engineering and Computer Science
[i] Corresponding Author: address to buttsc@uci.edu

August 21, 2023

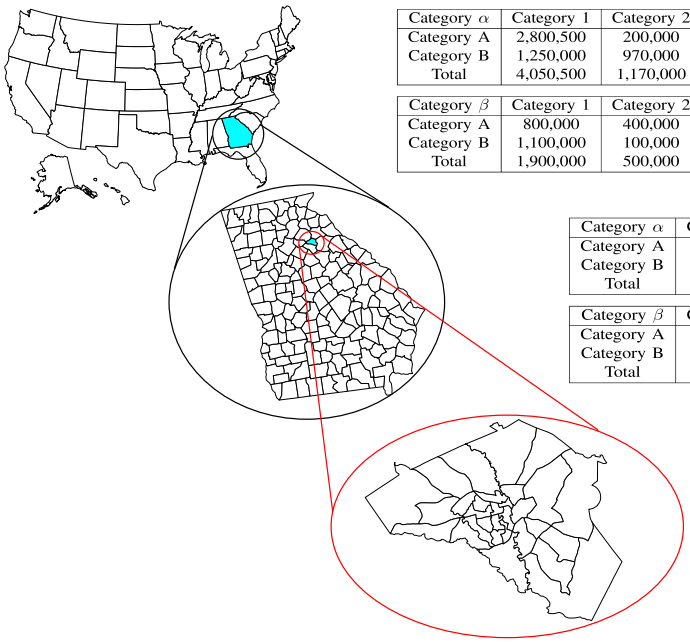
[1]This work was supported by NSF award SES-1826589.

Abstract

Geospatial population data is typically organized into nested hierarchies of areal units, in which each unit is a union of units at the next lower level. While there is increasing interest in analyses at fine geographical detail, these lowest rungs of the areal unit hierarchy are often incompletely tabulated due to cost, privacy, or other considerations. Here, we introduce a novel algorithm to impute crosstabs of up to three dimensions (e.g., race, ethnicity, and gender) from marginal data combined with data at higher levels of aggregation. Our method exactly preserves the observed fine-grained marginals, while approximating higher-order correlations observed in more complete higher-level data. We show how this approach can be used with U.S. Census data via a case study involving differences in exposure to crime across demographic groups, showing that the imputation process introduces very little error into downstream analysis, while depicting social process at the more fine-grained level.

1 Introduction

Many data sources, including the U.S. Census and organizations using Google's S2 projection system [1], provide geospatial population data organized into a nested hierarchy of areal units. In such hierarchical structures, each areal unit at a given level can be expressed as the union of a set of units at the level below, in turn being part of a single parent; each level is hence a spatial partition of the region of interest (Fig. 1). Many sociological questions involve the cross-tabulation of population properties within such units with other quantities (e.g., environmental, ecological, political, economic, or other variables that vary across regions). With the advent of increasingly well-developed spatial data sets (Rose et al., 2021; Facebook Connectivity Lab et al., 2016), performing such analyses at increasingly fine geographical resolution is of substantial interest (Thomas et al., 2020).



Category α	Category 1	Category 2	Total
Category A	2,800,500	200,000	3,000,500
Category B	1,250,000	970,000	2,220,000
Total	4,050,500	1,170,000	–

Category β	Category 1	Category 2	Total
Category A	800,000	400,000	1,200,000
Category B	1,100,000	100,000	2,220,000
Total	1,900,000	500,000	–

Category α	Category 1	Category 2	Total
Category A	100,000	15,000	115,000
Category B	120,000	80,000	200,000
Total	220,500	95,000	–

Category β	Category 1	Category 2	Total
Category A	75,000	45,000	120,000
Category B	120,000	15,000	135,000
Total	195,000	60,000	–

Category α	Category 1	Category 2	Total
Category A	?	?	213
Category B	?	?	104
Total	180	137	–

Category β	Category 1	Category 2	Total
Category A	?	?	115
Category B	?	?	200
Total	195	60	–

Figure 1: An example of a hierarchical data structure on areal units, using the US Census areal unit hierarchy. We are interested in tabulating population with respect to three hypothetical dimensions, represented by respective category sets $\{\alpha, \beta\}$, $\{1, 2\}$, and $\{A, B\}$. At higher levels in the hierarchy, we may have complete areal unit data with respect to all categories; but for small units, we may have only marginal information (third table). By combining marginal information at fine-grained units with associations observed in the more complete parent data, we impute cross-tabulations for the fine-grained units.

[1] The S2 geometries use nested areal units on the sphere, and can be used to describe spatial relationships across Earth based on sets of locations and attributes.

In practice, however, such fine-grained analyses can still encounter problems of data availability. For instance, while detailed Census data is publicly released at higher levels of the Census geography (e.g., counties), incomplete data is released at smaller geographical scales (e.g., blocks and block groups). This issue is not unique to the U.S. Census: releasing fully detailed information at fine scale poses challenges of acquisition cost (there are vastly more small areal units than large ones), availability (key variables may not be obtained at all scales), distribution and maintenance costs, and privacy considerations. Where information for smaller units is available, it is often available only marginally (i.e. summed across all values of a covariate), without the cross-tabulation needed to study many demographic processes. For instance, we may know how many individuals reside in a given unit by race, by ethnicity, and by gender, but we may not know how many White Hispanic women reside there. Raising our level of analysis to the smallest unit with complete tabulation may resolve this difficulty, but at cost of "blurring" spatial heterogeneity. Particularly when studying phenomena that occur on small scales - e.g., neighborhood interactions, exposure to crime or other events, or immediate access to local amenities - this causes problems for analysis.

While there is no perfect substitute for complete data, the presence of incompletely tabulated data suggests the viability of imputation strategies: even one-way marginals can be powerfully constraining, and two-way marginals even more so. When marginals can be combined with information on correlations from higher-order units with complete data, it may be possible to accurately estimate the local tabulation in a way that preserves all known quantities. This preserves spatial heterogeneity and permits fine-grained analysis, while also making use of more complete information where available. Surprisingly, this approach to the multi-way areal unit imputation problem appears to have been overlooked in prior literature, though we draw on a number of related developments in our work (as described below).

In this paper, we introduce a method for imputing cross-tabulated count data organized into a nested system of hierarchical bins, that is highly parallelizable and hence applicable to large systems (prominently including the U.S. Census). We focus on the case of data that is cross-tabulated with up to three different discrete features, each of which may take on a number of values (i.e., a three-way crosstab); our approach combines lower-order information on marginals from the focal bin with more complete, higher-order marginals from the bin's parent to impute the full multi-way array. We can verifiably preserve all available information on the focal bin (assuming that such data is consistent), while approximating higher-order information to the extent possible given low-order

constraints. Our technique also allows either point estimation, or simulation of draws from the conditional maximum entropy distribution of the target array given the observed data constraints, supporting use cases such as multiple imputation that is capable of offering consistent uncertainty measures (Rubin, 1996). As an illustration of the method, we apply our approach to imputation of small areal unit data using the 2010 U.S. Decennial census, demonstrating how it enables fine-grained ecological analysis (here, of differences in exposure to crime) despite data constraints.

2 Prior Work

As noted, the specific problem of marginal-preserving multi-way count-data imputation from combined marginal and hierarchical information seems not to have been addressed in prior work. However, a number of related problems have been studied, solutions to which inform our own approach. By way of background, we thus begin by reviewing related results on small areal unit estimation and imputation for three-way cross-tabs, both of which set the stage for our work.

2.1 Small Areal Unit Estimation

In its more general context, the problem of inferring characteristics of (usually small) geographical regions is known as the small area estimation (SAE) problem. This is a challenge that arises in many different fields, and work on SAE has likewise bridged a number of disciplines, including but not limited to sociology, demography (Morrison, 1971), and statistics (Graham et al., 2009; Bunea and Besag, 2000). While small areal unit estimation often deals with estimation of population demographics *per se*, some work goes beyond this to examine covariates such as poverty or disease (Pfeffermann et al., 2013; Molina and Rao, 2010). As noted in this paper, these techniques are also applicable to examinations of crime exposure in a population. There are numerous strategies for this problem, ranging from simpler strategies such as uniform imputation (completely uninformed at the small geographic unit) or spatial smoothing techniques such as kriging that attempt to flexibly exploit spatial autocorrelation across units (Bennett et al., 1984; Mooney et al., 2020), to more informed model-based approaches (Cohen and Di Zhang, 1988; Steinberg, 1979). Related to this work, some literature has specifically examined work on maintaining structural constraints and the use of model assisted approaches (Espuny-Pujol et al., 2018; Luna et al., 2015; Moretti and Whitworth, 2020).

In the field of criminology, the interest in estimating models in which crime is an outcome measure

in increasingly small geographic units has resulted in a need for small area estimation. Whereas some scholars have simply utilized a uniform imputation strategy to assign data from a larger geographic unit to a smaller unit, another strategy occasionally utilized synthetic estimation for ecological inference (Boessen and Hipp, 2015). This strategy requires the assumption that the relationships between variables in the larger geographic units are the same as the relationships within the smaller geographic subunits.

Surprisingly few studies in criminology or sociology have explored the exposure to crime of different demographic groups. Arguably, this state of affairs is due to the difficulty of obtaining crime data at a more granular scale. For example, one study measured the context of small suburban communities (defined as population less than 10,000) in assessing the exposure to crime Alba et al. (1994). Another study in Cleveland aggregated census tracts to "neighborhoods," and thus an even larger geographic unit (Logan and Stults, 1999). Yet another study measured the context as police precincts in New York City, which are larger yet given that there were 75 at the time of the study in a city with over 7 million residents (McNulty, 1999). The challenge is that such units may be too large to capture the environment of a specific person, or group of people. This issue arises in the context of other social exposures, as well; for instance, Thomas et al. (2020, 2022) provide evidence that both infection hazards and social exposure to others' morbidity and mortality in the early COVID-19 pandemic was affected by local variation in network structure influenced by housing and demographic factors at or below the block scale. Such differences in exposure may affect not only immediate health outcomes, but also responsiveness to public health interventions, with consequences for both policy effectiveness and health disparities.

The SAE problem is computationally challenging, both because it often involves discrete optimization (e.g., for population counts) and because SAE solutions are often intended to be used at scale: given that e.g., the U.S. has over 8 million Census blocks, and there are over 1.6 billion level 14 S2 cells worldwide (each about 500m across), efficiency can be a significant concern. As such, work on this problem has spurred a range of computational advances, from algorithms to actually perform estimation *per se* (Graham et al., 2009; Vermunt et al., 2008) to the evaluation of produced results (Pfeffermann and Correa, 2012). Many of the more statistically principled algorithms derive from the literature on hierarchical Bayesian modeling, which provides numerous conceptual and statistical tools for flexible estimation and incomplete pooling of information across units (King et al., 1999). While these frameworks often require significant computational resources, as the Markov

Chain Monte Carlo (MCMC) algorithms required for fitting and simulating draws from such models (Rosen et al., 2001) are very computationally intensive, the algorithms enable estimations of more complex problems where the joint probability functions are not in closed form.

In this context, our work contributes to the SAE literature by implementing an algorithm for Small Areal Unit Estimation that can produce imputed cross-classification data for areal units that satisfy a complex constraint structure (guaranteeing that imputations exactly preserve one and two-way marginal totals, and are integer-valued), while also including information from higher-order units. Our technique draws on the statistical strengths of the SAE literature by leveraging a hierarchical model, extending the work of (Bunea and Besag, 2000) by including additional information about the composition of larger areal units in the imputation process.

2.2 Imputation for Cross-tabulated Count Data

Apart from the SAE problem, our work is also related to the general problem of imputation for cross-tabulated count data. In general form, this problem involves a target matrix $N \in \mathbb{N}^{c_1 \times \cdots \times c_d}$ (with c_i being the size of the ith dimension, $\mathbb{N}$ the natural numbers, and $\times$ the Cartesian product), from which only a subset of cells (or, in many cases, marginal totals) is observed. The problem is then to produce a matrix $\hat{N}$ that approximates N, while preserving all observed quantities. For the purposes of this paper, we specifically focus on the three-way case (i.e., $d = 3$ in the above), as this case allows for significant variability for table interiors. Solving the three-way case also provides two way tables via marginal counts. Naively, the most basic option for three-way imputation is to evenly allocate population to each cell in the three-way crosstab. This preserves the 0-way marginal (i.e., the population total), but not other marginals. One-way or two-way marginals can be preserved by a continuous relaxation of the problem in which each cell is given the same value used as the expected value in the corresponding k-way Chi-Square test (McHugh, 2013), but this does not provide an integer solution - obtaining integer solutions that exactly satisfy the marginal constraints can be substantially more difficult (Bunea and Besag, 2000).

Beyond preserving marginal (or other) information in N, one may seek to preserve (or approximate) more general patterns of associations (e.g., correlations among category memberships). Again, the continuous relaxation of this problem is substantially simpler than the exact version, and indeed it has been extensively studied in the context of log-linear models (Clogg and Eliason, 1987). Log-linear models represent the expected count for each cell in an array as a multiplicative com-

bination of interactions, such that the log expectation has a linear form; expected marginals are easily preserved in this framework by incorporating parameters derived from observed marginals, but higher-order associations between category memberships can also be employed. Although the simplest approaches to inference for log-linear models are based on maximum likelihood estimation under the assumption that counts are conditionally Poisson distributed (exploiting the resulting exponential family structure), Bayesian and other forms of regularized inference (Graham et al., 2009; Vermunt et al., 2008) have also been employed. Log-linear models are thus powerful and flexible tools for obtaining conditional cell distributions that preserve *expected* patterns in a target matrix, though they do not solve the problem of preserving exact marginals.

Exact preservation of higher-order properties is more difficult, and generally requires specialized algorithms. In the context of graph construction (viewing a binary adjacency matrix as a two-dimensional matrix of 0 or 1 counts), a large literature has emerged on methods for preserving row/column marginals (i.e., degree sequences), as well as degree mixing and block marginals (i.e., mixing rates); see e.g. (Tillman et al., 2019) for a review of several common cases. Construction algorithms - which produce an instance $\hat{N}$ exactly satisfying some target properties of N are of somewhat limited value for imputation, as they make no guarantees that the arrays constructed are representative of the set of feasible solutions (and generally they are not). Fortunately, however, it is often possible to construct Markov chain Monte Carlo (MCMC) algorithms that, given a feasible instance of $\hat{N}$, will simulate draws from a uniform (or other) target distribution over the set of feasible imputations. For our purposes, the most relevant work is that of Bunea and Besag (2000), who provide an algorithm for sampling three-way count arrays that approximate a target distribution while preserving all two-way marginals. (When the two-way marginals are not available, Monte Carlo methods are available to construct data based only on one-way margins (Bunea and Besag, 2000), though we do not pursue this here.) We leverage and modify this procedure, using it to design an annealing algorithm that generates single imputations preserving both two-way marginals and higher-order correlations (a necessary goal for high-volume applications); in turn, we produce our target distributions using the log-linear modeling approach described above, exploiting the spatial hierarchy of areal unit data to obtain correlation information from higher-level units while preserving lower-level marginals.

The primary contribution of this paper is the implementation and development of a technique to impute three-way crosstab data that exactly preserves a set of integer marginals. Existing imputation

techniques (including many of the ones discussed in this section), have difficulty with this kind of constraint structure. We leverage work on existing imputation techniques that allows for the incorporation of higher order spatial data to improve the quality of the imputed data.

3 Technical Description

3.1 Data Representation

As discussed above, we are interested here in the specific case of imputing an unknown three-way array of counts, $n \in \mathbb{N}^{I \times J \times K}$, for which the two-way marginals (i.e., quantities of the form $n_{ij\cdot}, n_{i\cdot k}, n_{\cdot jk}$) are known. This array is assumed to represent the cross-tabulation of entities within a given areal unit, for which the corresponding cross-tabulation of entities within a parent unit n^H, is fully observed. Our goal will then be to impute $n|\{n^H, n_{ij\cdot}, n_{i\cdot k}, n_{\cdot jk} : i \in 1, \ldots, I, j \in 1, \ldots, J, k \in 1, \ldots, K\}$, while satisfying all observed marginals.

3.2 Imputation Method

To impute the data contained in the three-way marginal array, we extend the work of Bunea and Besag (2000). Using this algorithm as a baseline, we take a valid starting three-way array and use MCMC to simulate draws from the distribution of valid three-way arrays, given the set of two-way marginals that constrain it and a target distribution at a higher level of geography. We employ simulated annealing to both find the valid starting point and to find a maximum-probability array with respect to the target distribution, a robust heuristic optimization procedure that helps avoid becoming trapped in local maxima. More details on the imputation process can be found in Algorithm 1.

3.2.1 The Target Distribution

Our algorithm, which is discussed in Section 3.2.6 requires a target distribution to be approximated (subject to our marginal constraints); since the two-way constraints will automatically account for all known information about the target array (n), the role of this distribution is to provide information regarding three-way associations that cannot be obtained for the target areal unit. We here employ the conditional log-linear model for the fully observed contingency table of the parent of the focal

areal unit, n^H, to generate the target distribution. As a log-linear model is a discrete exponential family on the space of count arrays, it can be understood as leading to the maximum entropy distribution on the space of such arrays given the observed statistics and appropriate choice of reference measure (Darroch and Ratcliff, 1972; Jaynes, 1982). Concretely, when applied to statistics based on table margins, it results in an inferred distribution that preserves the expected margins in the contingency table, while maximizing the uncertainty of the cell values given those expectations. Here, we base our target distribution on the three-way effects observed in n^H, while simulating conditional on the two-way margins of n; this gives a maximum-entropy approximation to the three-way structure of n^H, net of the (exactly preserved) marginal constraints of n, which allows us to use information from higher-order areal units to inform imputation for low-order units. This is accomplished as follows.

A saturated log-linear model contains sufficient statistics of effects at different levels in the contingency table. Specifically, for an array defined by three dimensions/covariates i, j, k, we can specify the model as:

$$\mathbf{E}(n_{ijk}) = \tau\tau_i\tau_j\tau_k\tau_{ij}\tau_{ik}\tau_{jk}\tau_{ijk}$$

where $\mathbf{E}(n_{ijk})$ denotes the expected count of the i, j, k cell; τ is the intercept, or the main effect of the contingency table; τ_i, τ_j, τ_k denote the marginal effects for the dimensions i, j, k, respectively; $\tau_{ij}, \tau_{ik}, \tau_{jk}$ denote the two-way interaction effects (with dimensions as above); and, finally, τ_{ijk} denotes the three-way interaction effect over all three dimensions. (Note that fixing the expectation, when combined with the assumption of a maximum entropy distribution over the set of possible matrices under a Poissonian reference measure,[2] fully specifies the model.)

With information of two-way margins available for the target areal unit, one could estimate the marginal effects, and the two-way interaction effects. However, this is not sufficient to provide information about the three-way interaction term. Here, we approximate the three-way interaction effect for the contingency table of our target areal unit by the effect observed for its parent areal unit (treating the former *de facto* as a sample from the latter). This can also be viewed as a two-step process, where we first get an expected cell input based on information at the lower-level, and then re-calibrate it using information of the three-way interaction effect from the higher level. Formally, take $\mathbf{E}^L$ to be the expectation given all observable margins of the lower-level (i.e., target) areal unit;

[2] I.e., $h(x) = \prod_i (x_i!)^{-1}$, where the product is over cells. This amounts to assuming indistinguishability of individuals within groups.

then we have

$$\mathbf{E}^L(n_{ijk}) = \tau^L \tau_i^L \tau_j^L \tau_k^L \tau_{ij}^L \tau_{ik}^L \tau_{jk}^L, \tag{1}$$

where $\tau_{\cdot}^L$ reflects parameter estimates based on the marginals of the observed (lower-level) areal unit. Now, letting τ_{ijk}^H be the estimate of the three-way effect from n^H, we employ the specification

$$\mathbf{E}(n_{ijk}) = \mathbf{E}^L(n_{ijk}) \tau_{ijk}^H. \tag{2}$$

Thus, we employ data from n^H to fill in the "missing piece" that cannot be obtained from n itself, while retaining all lower-order information from n.

Owing to the exponential family properties of the log-linear model, the parameters τ are easily estimated from the observed counts. The parameters describe the ratio between expected cell counts with and without the effects they represent; therefore, they are equal to 1 when absent. The general effect τ is equal to the grand mean of the contingency table, i.e. $\overline{n_{\cdots}}$. The one-way marginal effects are in turn equal to the ratios between the corresponding marginal means and the grand mean. Formally,

$$\tau_i = \frac{\overline{n_{i\cdot\cdot}}}{\overline{n_{\cdots}}}$$
$$\tau_j = \frac{\overline{n_{\cdot j\cdot}}}{\overline{n_{\cdots}}}$$
$$\tau_k = \frac{\overline{n_{\cdot\cdot k}}}{\overline{n_{\cdots}}}$$

where $\overline{n_{i\cdot\cdot}}$, $\overline{n_{\cdot j\cdot}}$, and $\overline{n_{\cdot\cdot k}}$ denote respective marginal means. The two-way interaction effects, in turn, are equal to the ratios of the respective two-way means to the expectations of those means arising from the respective one-way means. Formally,

$$\tau_{ij} = \frac{\overline{n_{ij\cdot}}}{\frac{\overline{n_{i\cdot\cdot}}\,\overline{n_{\cdot j\cdot}}}{\overline{n_{\cdots}}}} = \frac{\overline{n_{ij\cdot}}\,\overline{n_{\cdots}}}{\overline{n_{i\cdot\cdot}}\,\overline{n_{\cdot j\cdot}}}$$
$$\tau_{ik} = \frac{\overline{n_{i\cdot k}}\,\overline{n_{\cdots}}}{\overline{n_{i\cdot\cdot}}\,\overline{n_{\cdot\cdot k}}}$$
$$\tau_{jk} = \frac{\overline{n_{\cdot jk}}\,\overline{n_{\cdots}}}{\overline{n_{\cdot j\cdot}}\,\overline{n_{\cdot\cdot k}}}$$

where $\overline{n_{ij\cdot}}$, $\overline{n_{i\cdot k}}$, $\overline{n_{\cdot jk}}$ denote the respective two-way means. Therefore, we may rewrite equation 1

in terms of observed counts as

$$\mathbf{E}^L(n_{ijk}) = \frac{\overline{n_{ij.}}\,\overline{n_{i.k}}\,\overline{n_{.jk}}\,\overline{n_{...}}}{\overline{n_{i..}}\,\overline{n_{.j.}}\,\overline{n_{..k}}}, \tag{3}$$

a quantity that is easily calculated.

With the first factor in hand, we now require only τ^H_{ijk}. As with the previous cases, the three-way interaction effect is equal to the ratio of the three-way marginal mean (here, identically the count of the i, j, k cell) to the expectation given the lower order effects.

Next, we process the three-way interaction effects using information from the higher-level unit. Similar to the previous derivations, the three-way interaction effect equals to the ratio of the cell with three-way interaction effect over that without the effect. Bearing in mind that all relevant counts here are for n^H, we have

$$\tau^H_{ijk} = \frac{\overline{n^H_{ijk}}}{\frac{\overline{n^H_{ij.}}\,\overline{n^H_{i.k}}\,\overline{n^H_{.jk}}\,\overline{n^H_{...}}}{\overline{n^H_{i..}}\,\overline{n^H_{.j.}}\,\overline{n^H_{..k}}}} = \frac{\overline{n^H_{i..}}\,\overline{n^H_{.j.}}\,\overline{n^H_{..k}}\,\overline{n^H_{ijk}}}{\overline{n_{ij.}}\,\overline{n_{i.k}}\,\overline{n_{.jk}}\,\overline{n_{...}}}, \tag{4}$$

which is again easily calculated from the observed arrays. In passing, we note that this expression for τ^H_{ijk} makes clear that it is already "normalized" with respect to the lower-order marginals of n^H; thus, differences between n and n^H in such quantities do not prevent τ^H_{ijk} from being used to model n (and, indeed, the three-way effects by construction do not affect any lower-order marginal expectations).

Putting these pieces together, the final target distribution is proportional to a product of Poisson distributions (a form that arises from the maximum entropy construction), whose expectations are functions of data from the target areal unit and its parent. The final target expectation for a given cell is the product of the expected distribution given the lower level information (Eq. 3), and the three-way interaction effect from n^H (Eq. 4), i.e.

$$\mathbf{E}(n_{ijk}) = \frac{\overline{n_{ij.}}\,\overline{n_{i.k}}\,\overline{n_{.jk}}\,\overline{n_{...}}}{\overline{n_{i..}}\,\overline{n_{.j.}}\,\overline{n_{..k}}} \cdot \frac{\overline{n^H_{i..}}\,\overline{n^H_{.j.}}\,\overline{n^H_{..k}}\,\overline{n^H_{ijk}}}{\overline{n_{ij.}}\,\overline{n_{i.k}}\,\overline{n_{.jk}}\,\overline{n_{...}}}. \tag{5}$$

3.2.2 Imputating a Three-Way Array

While the target distribution that we specify above will be used to find a maximum probability array, any three-way array imputed must match observed two-way marginals. Thus, we separate our imputation procedure into two distinct steps. First, we construct an array that satisfies the con-

straints imposed by the two-way marginals. Then, with this array that matches observed marginals, we optimize the array with respect to the target distribution, preserving two-way marginals. Each of these components is non-trivial. Due to the integer constraints, finding an array that matches observed marginals (a valid array) is not possible using standard techniques (such as the expected count array formed when performing a generalized Chi-Square test). Likewise, for the three-way case, optimizing the array to maximize a target distribution is a challenging task.

3.2.3 Constructing a Valid Array

Our algorithm to impute a target three-way array begins by finding an array that satisfies the observed two-way marginals. This component must solve an array *construction* problem, prior to the optimization problem discussed in the second part of the algorithm (see section 3.2.6). This part of the algorithm is concerned only with satisfying the two-way and integer constraints, and does not consider the target distribution for array construction.

Our strategy (detailed with pseudocode in Algorithm 1) can be broadly described as follows. Algorithm 1 also includes optimizations discussed in Section 3.2.4. For a full description of the algorithm, see section 3.2.5. Our algorithm initializes an array using data from the zero-way marginal (i.e. the total array population). All population is divided equally across the array, with any remainder allocated to the first cell. This is detailed in lines 1 and 2 of Algorithm 1. This initial state ensures that the total population of the array and the integer constraints are satisfied. However, it is unlikely that this initial array state will satisfy the constraints imposed by the one or two-way marginal values. We can define the deviation of our constructed array and the observed two-way marginals with the sum of the absolute values of the differences between the two-way marginals of our constructed array and the two-way marginals of the target array. We then seek to minimize this deviation.

We utilize a strategy of *simulated annealing* to produce a valid array from the initial state of our constructed array. Simulated annealing is a heuristic optimization technique designed to find the global minimum of an objective function, with minimal assumptions regarding the function and search space. This strategy will simulate moving values (individuals) between cells in the array, keeping track of the deviations between the simulated marginals values and the target marginal values. A single move will decrease the value of one cell and increase the value of another cell. However, the array is not considered as a valid state unless all cells in the array are non-negative. If

there is a negative value in the array after a proposed move, we draw a new move based on the state of the proposed array. This process of drawing proposed changes will continue until a valid state of the array is drawn. This valid state will be proposed as a new state of our constructed array, and we compute the marginals of this array, as well as the arrays deviation from the observed marginals.

The annealer will always accept moves in the array that decrease the deviation between simulated and target marginals, as these moves will bring the simulated array closer to one that satisfies the constraints of the two-way marginals. The annealer will also accept moves that *increase* the deviation with a probability equal to $\exp(\frac{D_C - D_P}{T})$, where D_C is the deviation between marginals for the current array state, D_P is the deviation from the marginals for a proposed array state, and T is a temperature parameter. We still accept moves that increase deviation from the target marginals in order to prevent the annealer from finding a local minimum in the error space. However, the temperature parameter T will scale the likelihood that disadvantageous moves are taken. At high temperatures, accepting moves that increase our deviation is more likely, while lower temperatures make it much more difficult to accept these moves. The idea behind simulated annealing is to begin with a high temperature and allow the state of the array to vary more easily with respect to our deviation. This will help to prevent the state of our array from being stuck in a local minimum of deviation. As the annealer runs, we decrease the temperature geometrically, which will minimize the deviation by the end of the annealing run. Although convergence was easily obtained in the cases studied here, it should be noted that it is possible that the annealer will not converge to a a valid array. In this case, repeatedly restarting with a higher temperature and using a slower cooling schedule until convergence is obtained is a practical strategy. It should be noted that, regardless, convergence is always *verifiable*, since we can always determine whether or not our current array satisfies the target marginals (and, if not, the degree of divergence).

While the annealer we describe here should produce a valid array that matches the constraints from the two-way marginals, the basic version of the algorithm requires us to recompute the two-way marginals every time we get a new state for our constructed array. While computing the marginals of the array once does not take a significant amount of time, computing them for every array state does add a high cost in computational time. To avoid this cost and improve the algorithm runtime, we introduce several optimizations, detailed in section 3.2.4.

3.2.4 Optimizations for the Construction Algorithm

The algorithm detailed in section 3.2.3 provides an array that will satisfy both the two-way marginal and integer constraints. However, due to the requirement to recompute the two-way marginals of proposed arrays, the algorithm can be expensive. To achieve better performance for larger datasets, we in practice implement a version of the algorithm that uses a change score. Specifically, we compute the *difference* between the initial state of the array and the target marginals, keeping track of these persistent errors. When we move values between array cells, we then update this persistent error by subtracting a person from the departure cell of the marginals, and adding a person to the arrival cell of the marginals (rather than recalculating the marginals anew). This persistent error is equivalent to using the error metric from section 3.2.3, but does not require recomputing marginals.

The updated error metric will avoid the computational cost of recalculating the marginals, but does require additional components. As noted above, we need to compute a map between the three-way array and each of the two-way marginals. This mapping will allow us to remove a person from the relevant cell of the two-way marginals and add them to the arrival cell when making a move in the three-way array. However, we only compute this mapping once, and can then refer to it when making moves in the three-way array. In Algorithm 1, the helper function mapIndexToMarginal will take a three-way array index and map it to the X,Y, or Z marginals respectively.

3.2.5 Description of Construction Algorithm

Algorithm 1 provides pseudocode for the construction of a three-way array that matches integer and two-way marginal constraints. This algorithm uses a set of two-way marginals X, Y, and Z. The name of the marginal refers to the direction that we sum across the array to produce each marginal. We also need a set of helper functions for this algorithm. The functions xMargins(a), yMargins(a), and zMargins(a) each take a three-way array and produce a a two-way marginal. The function RandomInt(a,b) produces a random integer from a to b, inclusive. mapIndexToMarginal(a) takes a three-way array index and maps this index to a two-way marginal index. We use one of these functions for the X,Y, and Z marginals. Finally, numNegative(a) takes a three-way array and returns the number of negative values in the array.

Lines 1 and 2 of 1 produce the initial state of the array. The variable $numTotalCategories$ provides the number of cells in the array. Next, lines 3-5 produce the deviation of the initial state

of the array from the observed marginals. These values will be used to compute an error metric, and will be used as a persistent deviation throughout the algorithm. Next, we specify how many arrays we will simulate with the annealer (M), and begin simulating arrays within the loop on line 6. Lines 7-10 are used to initialize the state of a proposed array, as well as the deviation that this proposed array would have from the target marginals.

After the deviation and array values have been initialized for our simulation, the second while loop (on line 11) begins the search for a valid array state to compare to our initial array state. Lines 12-16 draw two three-way array indices, i and j, and ensures that they are different. We also produce the corresponding two-way array indices for X,Y, and Z using lines 17-22, which use the mapIndexToMarginal() helper function, which uses a pre-computed map between three-way indices and two-way marginal indices. After we produce all of the necessary indices, we move a value in the three-way array from index i to index j, which simply adds one to the j^{th} cell of the array, and subtracts one from the i^{th} cell. Lines 25-30 also keep track of the move in the three-way array in the two-way marginals.

For this algorithm, a valid state of the array is one in which all cells are non-negative. At the end of every proposed move, we check to see if this condition is met (Lines 31-33), and if so, immediately end the search for a new move. If any array cells are negative, we draw another move and continue until we have a non-negative array.

With our non-negative array, we can next compute an error metric that uses the deviations of the initial array and the deviation of the proposed array. Line 35 computes this error, which is the difference in the absolute values of the deviation summed for the initial array and the proposed array. This value would be positive if the proposed array reduces the deviation from the target marginals, while it would be negative if the proposed array increases the deviation. The probability that our proposed array is accepted and becomes the current state of the array is computed on line 36, and is simply the error term divided by the temperature term, exponentiated. Lines 37-42 check to see if we accept the proposed array state. If we do, then the proposed array becomes the current array. Likewise, the deviations from our proposed array state would become the initial deviations for the next iteration of the loop. If the array state is not accepted, the proposal is discarded and we begin from the initial state again.

Before the loop iteration ends, line 44 cools the temperature parameter. We use a geometric cooling schedule, where the temperature parameter is multiplied by a constant for each run of the

array. The final state of the array is returned on line 46, and should have minimized the difference between the marginals of the array and the target array's marginals. In the implemented algorithm, we also check to see if the deviation between marginals is zero, and end the annealing process if it is. It is important to note that while this description has assumed that we are using this algorithm for single imputation, the algorithm will also work for multiple imputation. By fixing the cooling schedule and temperature parameters at 1, we are able to draw directly from the target distribution, which will be produced by the Markov chain.

3.2.6 MCMC Optimization Algorithm

The construction algorithm detailed above will produce a three-way array that satisfies the integer and two-way marginal constraints. The second component of our algorithm optimizes the three-way array's values with respect to a target distribution, using the specification described in Section 3.2.1. Our algorithm builds on the approach described by Bunea and Besag for simulating from three-way count arrays (Bunea and Besag, 2000). This algorithm assumes that we start with a legal imputation of the target array (i.e. an array that satisfies the set of two-way marginals and has no negative values), as well as a target distribution.

Both our algorithm and the Bunea and Besag algorithm use MCMC to simulate three-way array configurations. The transitions between states use a *basic move*, in which a person is moved from one cell to another in the array. However, as each state of the array will is required to match the two-way marginal constraints, eight cells in the three-way array are modified in total for each basic move. Additionally, we can define the log-likelihood of a given array state under the target distribution, which we will denote $l(n)$, where n is a three-way array. For two arrays, a current and proposed array, the probability for the Markov chain to accept the proposed move is $\exp(l(n') - l(n))$, where n is the current array state, and n' is the proposed array state. Because the basic move preserves all lower-order marginals, validity of an array produced by this method is guaranteed (i.e., any array generated from a valid starting point will always be a valid array).

We modify Bunea and Besag's algorithm by using simulated annealing. In the construction component of this algorithm, we used annealing to optimize the state of an array to minimize deviation from the observed marginals. As our goal in this part of the algorithm is to find the most likely configuration of the algorithm under the target distribution, we can use simulated annealing for better single imputation quality. The addition of simulated annealing is straightforward, and only requires

Algorithm 1 Produce a three-way array that satisfies a set of two way marginals X, Y, Z

$n_{ijk} \leftarrow$ floor(sum(X) / $numTotalCategories$)
$n[1] \leftarrow n[1]$+sum(X)−floor(sum(X)/$numTotalCategories$)
$xError \leftarrow$xMargin(n) − X
$yError \leftarrow$yMargin(n) − Y
$zError \leftarrow$zMargin(n) − Z
while $M > 0$ **do**
  $n' \leftarrow n$
  $xError' \leftarrow xError$
  $yError' \leftarrow yError$
  $zError' \leftarrow zError$
  #Ensure that the next state of the array has no negative values
  while $K > 0$ **do**
    $i \leftarrow$RandomInt(1, $numTotalCategories$)
    $j \leftarrow$RandomInt(1, $numTotalCategories$)
    while $i == j$ **do**
      $j \leftarrow$RandomInt(1, $numTotalCategories$)
    end while
    # Map the origin and destination of the move to the marginals
    $mappedI_X \leftarrow$mapIndexToMarginalX(i)
    $mappedI_Y \leftarrow$mapIndexToMarginalY(i)
    $mappedI_Z \leftarrow$mapIndexToMarginalZ(i)
    $mappedJ_X \leftarrow$mapIndexToMarginalX(j)
    $mappedJ_Y \leftarrow$mapIndexToMarginalY(j)
    $mappedJ_Z \leftarrow$mapIndexToMarginalZ(j)
    #Do the move in the three-way array
    $n' \leftarrow$moveAPerson(n, i, j)
    $K- = 1$
    # Update the marginal deviations
    $xError'[mappedI_X] \leftarrow xError'[mappedI_X] - 1$
    $yError'[mappedI_Y] \leftarrow yError'[mappedI_Y] - 1$
    $zError'[mappedI_Z] \leftarrow zError'[mappedI_Z] - 1$
    $xError'[mappedJ_X] \leftarrow xError'[mappedJ_X] + 1$
    $yError'[mappedJ_Y] \leftarrow yError'[mappedJ_Y] + 1$
    $zError'[mappedJ_Z] \leftarrow zError'[mappedJ_Z] + 1$
    #If our array is non-negative, end the search for a move
    if $numNegative(n') == 0$ **then**
      break
    end if
  end while
  #Evaluate the relative error of our proposal and current arrays
  $error \leftarrow$sum(abs($xError$),abs($yError$),abs($zError$))−sum(abs($xError'$),abs($yError'$),abs($zError'$))
  $transitionProbability \leftarrow \exp(error/temperature)$
  if uniform(0, 1) < $transitionProbability$ **then**
    $n \leftarrow n'$
    $xError \leftarrow xError'$
    $yError \leftarrow yError'$
    $zError \leftarrow zError'$
  end if
  $M- = 1$
  #Cool the chain
  $temperature \leftarrow temperature * 0.9$
end while
$return(n)$

us to modify the acceptance probability with a temperature parameter. As with the construction algorithm, the temperature parameter will scale the probability that the optimization algorithm accepts proposed arrays that are less likely under the target distribution. Higher temperature values (above 1) will increase the likelihood that less likely array configurations are accepted. Likewise, as temperatures approach zero, the probability of accepting lower-probability array configurations also goes to zero. A fixed temperature at 1 will accept new array states by exactly evaluating the likelihood within the target distribution. A benefit of updating this algorithm to use simulated annealing is that the algorithm can be run in both single and multiple imputation modes. In single imputation mode, the temperature would be set above 1, and the cooling schedule would be set below 1. However, as mentioned above, by setting the temperature and cooling schedule to 1, we would draw directly from the target distribution, which enables multiple imputation.

3.2.7 Optimization Algorithm Description

Algorithm 2 provides pseudocode for the optimization of a three-way array with respect to a target distribution corresponding to the maximum entropy distribution on n conditional on $\mathbf{E}(n)$ and the two-way marginals. We need several helper functions for this implementation. First, we use a helper function doBasicMove(n), which takes a three-way array as an input, and moves someone from one cell in the array to another, maintaining all two way marginals. As discussed above, this basic move modifies eight cells of the three-way array, and is described in more detail in Bunea and Besag (2000). We also use the helper function numNegative(n), which takes a three-way array as an input, and outputs the number of negative values in the array. Finally, the helper function numNegativeOne(n) takes a three-way array as an input, and oututs the number of negative ones in that array.

The implementation of our algorithm relies on a Markov chain that cools as the annealer runs. The total length of the Markov chain is $M * L$, where M is the number of times we cool the Markov chain, and L is the number of iterations we run the Markov chain at each temperature. At each step of the Markov Chain, we start by proposing a basic move on the target three-way array n (Line 3). n', the proposed three-way array, will match all two-way marginal and integer constraints. Next, we follow the Bunea and Besag algorithm by checking the number of negative values (specifically negative ones) in n'. If there is exactly one negative one in the array, we draw a new basic move from n', and continue to do so until either there is either more than one negative value in n', or n' becomes a non-negative array. If n' is an array with more than one negative value, we discard the

proposal, keeping the original state of the array. However, if n' is a non-negative array, we compute the ratio of the likelihoods for n' and n (in log space) under the target distribution, and divide this log-likelihood by a temperature parameter (Line 11). When exponentiated, this is the probability that the proposed array is accepted as the next state of the Markov chain.

Every L iterations of our Markov chain, we cool the chain. Like the construction algorithm, we use a geometric cooling schedule for this algorithm, multiplying the temperature of the annealer by a constant every L iterations. As the temperature of the annealer decreases, the Markov chain will accept proposed states that are lower likelihood under the target distribution less often than at higher temperatures.

Algorithm 2 Impute a three-way crosstab

Require: target expectations $\mathbf{E}(n)$, array state n, initial temperature T, decay parameter $c = 0.94$

1: **while** $M > 0$ **do**
2: **while** $L > 0$ **do**
3: $n' \leftarrow$doBasicMove(n)
4: **while** numNegativeOne$(n') = 1$ **do**
5: $n' \leftarrow$doBasicMove(n')
6: **end while**
7: **if** numNegative$(n') > 1$ **then**
8: next
9: **else** {numNegative$(n') = 0$}
10: $randomNum \leftarrow$Uniform$(0, 1)$
 #Accept the proposed array with probability equal to the ratio of probabilities of proposed:current arrays
11: **if** $randomNum <$min$(1, \exp((\text{l}(n') - \text{l}(n))/T))$ **then**
12: $n \leftarrow n'$
13: **end if**
14: **end if**
15: $L- = 1$
16: **end while**
17: $M- = 1$
 #Cool the chain
18: $T \leftarrow T * c$
19: **end while**
20: $return(n)$

4 Validation of Imputed Data Quality

Above, we have provided algorithms for construction and imputation of three-way count arrays with targeted characteristics. Next, we describe the test imputations and metrics that we use to validate the quality of population data imputed using this approach. Our validation tests employ U.S. census

data on population distributions, using several levels of geographic aggregation. We also use two validation metrics to determine data quality.

4.1 Data used for Validation Runs

We use data from the 2010 U.S. census to assess the quality of our imputation technique. The U.S. Decennial Census published complete three-way population distributions at several levels of geographic aggregation. The Census uses a geographic hierarchy for their data products, with Census blocks aggregating into Census tracts, which themselves aggregate into counties.[3] We consider the three-way distribution of race, gender, and ethnicity within each geographically defined subpopulation (i.e. count data for each three-way category). Ethnicity has two categories, Non-Hispanic and Hispanic. Gender also uses two categories, Male and Female, while Race has 7 categories[4]. Given the national scale of the Census, these distributions provide a large dataset for us to test our imputation.

We perform two imputation studies in order to validate the approach. The first employs U.S. Census data across the entire United States, specified at the county and tract level. Here, target distributions are defined using three-way distributions at the county level. At the tract level, we use the two-way observed marginals for our target array. Full three-way distributions are publicly available at the tract level, which allows us to directly validate the quality of the imputation. We impute the three-way distribution of race, ethnicity, and gender for each of the 73,057 tracts in the United States. We perform both single imputation and multiple imputation for each tract, comparing true values against imputed counts.

Our second imputation study involves analysis of a social outcome (exposure to crime), using data specified at the tract and block levels. We specify a target distribution using full, three-way arrays available at the tract level, and use marginal two-way arrays that were published at the Census block level of aggregation. We chose to only impute the three-way arrays for one U.S. state (California), which contains 710,154 census blocks. Comparison of analysis at the tract level on actual versus imputed data provides another check on imputation quality. Although we cannot directly validate block-level imputation (since the three-way marginals are no available), we employ this for an illustrative case study described in section 4.3.2.

[3]There is also an intermediate level of aggregation known as the block group, but since their data availability is similar to blocks, we do not consider them here.

[4]Those categories being White, Black, Asian, Native American, Pacific Islander, Multiple Races, and Other

4.2 Imputation Parameters

For both of the validation samples, we use the same settings for both array construction and optimization. Additionally, the imputation calculations were completed on the same machine, using the same computing resources (facilitating timing comparisons).

First, we detail the parameters used for construction of a valid array. For each array we construct, we simulate up to 1,000,000 array states (M in Algorithm 1). We also allow for up to 1000 moves to find a new valid array state (K in Algorithm 1). We initialize our temperature parameter T to be ten times the error rate (deviation from the marginals) produced by the initial state of the array. In practice, all arrays produced by the implementation of Algorithm 1 were found to match the two-way and integer constraints, indicating that these parameters are sufficient for the heuristic optimization to succeed in finding valid array states.

Next, in the optimization component of the algorithm, we use a Markov chain of length 50,000 for each array. We cool this chain every 1,000 iterations, for a total of 50 annealing steps. The initial temperature parameter T is set to 10, with a cooling parameter c of 0.94. This allows the annealer to accept less likely array states more readily for half of the Markov chain, with the second half of the Markov chain behaving more strictly as a hill climber, seeking the maximum likelihood array state.

When doing multiple imputation, we fix both the temperature and the cooling parameter at 1 (i.e., we fix the algorithm at the target distribution, with no cooling). This allows us to draw directly from the distribution of array states that is specified by the target distribution and the marginal constraints. We use a thinning parameter of 1000 and a burn-in parameter of 1000, which were found to be adequate for convergence. In seeding the Markov chains for the multiple imputation draws, we initialize the optimization portion of our method with a draw from the single imputation mode of the algorithm. We do this to ensure that the Markov chain will burn-in, by starting it at a mode of the target distribution.

Each of our imputation studies was performed on an Intel Xeon E5-2599 V4 CPU. As the three-way array that is present for any areal unit does not depend on any other areal unit, this problem is trivially parallelizable. Thus, we used 30 cores for each imputation. We also introduce several special cases where we can directly solve the state of the three-way array. The first case is the one where there is no population in the array. Second, we can directly solve the "one-hot" case, or

arrays in which there is only one cell with any population. We can solve these arrays using only information from the two-way marginals, trivially imputing the array.

4.3 Metrics for Assessing Data Quality

We use two main methods for the assessment of data quality. Both of these metrics require observed data as a baseline. Therefore, we rely on the tract level imputation that is described above, as the full three-way tract arrays are published in addition to the two-way marginal data. Our first metric for assessing data quality relies on an error metric, and can be assessed on an individual array basis. We also provide a metric for assessing quality that depends on stable performance in a downstream analysis.

4.3.1 Error-Based Metric for Data Quality

The first metric we use to asses data quality measures the degree that an imputed three-way array departs from its observed values. In other words, this accuracy metric measures how many people are mismatched between a simulated and observed array. Our error metric, E is defined as $E = \sum_i |O_i - I_i|$, where O is the observed array for an areal unit, I is the imputed array for the same areal unit, and the sum is over entries of the array. This error metric simply represents the number of people who are misallocated by the imputation.

For purposes of expressing this error metric in a standardized manner, we divide E by the number of people present in the areal unit, which normalizes the error values to a range between 0 and 1. (In tracts where there is zero population, we define the metric to be zero.) This value is referred to as E_R, and describes the percentage of the tract that has been misallocated. Low values of this error metric indicate high quality imputed data.

We compare the errors produced by our algorithm to error rates produced by several other approaches. The first alternative to which we compare is the one described by Bunea and Besag, in which there is no simulated annealing, and we simply use a Metropolis algorithm to take a single draw from the distribution defined at the higher level of geography. We would expect that in this case, error rates would be broadly similar - since, if the Markov chain is burned in correctly, a random draw from that distribution is relatively likely to be from a high-probability region - but with higher excursions due to the fact that the algorithm will occasionally select plausible but low-probability arrays. This provides a point of comparison for the annealing algorithm, which uses the same target

distribution but attempts provide a maximum-probability array.

Next, we examine the case where we use the expected values provided by the target distribution as the final imputed values. These expected counts are produced using the log-linear framework described in section 3.2.1, which incorporates two way data from the target level of geography with three-way patterns present at the higher level of geography. Because the log-linear model is not constrained to satisfy integer constraints, it is expected to accumulate numerous errors; however, it nevertheless incorporates distributional information, and (being easy to compute) is an obvious practical alternative.

Finally, to examine the improvement produced by simulating the distribution of possible three-way states under the target distribution, we also examine the error rates when using the array generated by the construction algorithm as the final imputed value. This array will be "valid," in that it satisfies both integer constraints and the known two-way marginals, but not otherwise adjusted. We specifically examine this case to better understand how the space of three-way arrays may be constrained by the two-way marginals. This technique would also omit all data from the higher level of geography, so we can examine how only incorporating the local demographic effects (i.e. the two-way marginal constraints) may produce different arrays from the observed data.

4.3.2 Case Study for Quality Checking

While direct error assessment is the most natural way to evaluate imputation quality, it does not speak directly to downstream impacts on subsequent analysis: relatively poor imputation may in some cases prove adequate when downstream analyses are robust, while sensitive analyses may require very high degrees of imputation accuracy. While such sensitivity inevitably depends on the analysis involved, we here use a case study involving a spatially heterogeneous outcome - exposure to crime in one's vicinity - as a plausible example of how errors may or may not impact substantive conclusions. Specifically, we carry out our analysis at the tract level using both observed and imputed data, allowing us to compare results obtained in the two cases. For this purpose, we employ both single and multiple imputation, allowing us to compare the performance of both estimators at recovering observed-data results. Finally, as an illustrative procedure, we repeat our data analysis at the block level. Although not suitable for validation (since we do not have block-level observations), this analysis provides an example of how the imputation approach might be used in a realistic case, and how pushing analysis to finer levels of geographic detail can potentially impact our substantive

conclusions.

Our case study examines how exposure to crime near one's home is related to one's demographic characteristics. Crime is heterogeneously distributed, making members of some groups more likely than others to be exposed; such exposure may, in turn, feed concerns about neighborhood safety, willingness to access local affordances, and stress. To examine this association, we use crime data obtained from police agencies for the Southern California Crime Study (SCCS). In the SCCS, the researchers made an effort to contact each police agency in the Southern California region and request address-level incident crime data for six part 1 Uniform Crime Report (UCR) categories: homicide, aggravated assault, robbery, burglary, motor vehicle theft, and larceny. These data come from crime reports officially coded and reported by the police departments and provide locations of crime incidents around 2010 covering about 83% of the population in a five-county area (Los Angeles, Orange, Riverside, San Bernardino, and San Diego). Crime events were geocoded for each city separately to latitude/longitude point locations using ArcGIS 10.2, and subsequently aggregated to various units such as blocks and tracts. The average geocoding match rate was 97.2% across the cities, with the lowest value at 91.4%. These data have been used in several prior studies (Kubrin and Hipp, 2016; Kubrin et al., 2018). We specifically use the number of violent crime events that take place (homicide, aggravated assault, robbery), and compute the average over three years (2009-11) to smooth year to year fluctuations. Prior literature shows that one's exposure to crime is affected by many demographic features, including the ones that we have imputed in this paper (Alba et al., 1994; Logan and Stults, 1999; McNulty, 1999). The actual form of this relationship has not been particularly closely examined, however - particularly at the level of small areal units, which are needed to avoid averaging across areas with different crime rates. Thus, using both observed and imputed data at the census tract level (see section 7.1 for details on the imputation), we specify a saturated linear regression model (i.e. main effects, two-way interaction terms, and three-way interaction terms). Additionally, we specify the same models at the census block level, only using imputed data. We examine the block level model to compare whether the effects are similar to those of the tract-level model. At the tract level the mean number of crime events in the data is 42.38 events, with a minimum of zero events and a maximum of 666 events. At the block level, we use an additional buffer around each areal unit. This buffer has a radius of 1km. The mean number of crime events for the block level data is 112.22 events, with a minimum of zero events and a maximum of 2234 events.

The census geographies that we use here are adjacent levels of the census spatial hierarchy. Census tracts compose counties, and are often relatively large. For tracts represented in the SCCS, the average tract population in 2010 is 4604 people. While tracts can provide an overview of population distributions across space, the census block level is much more granular (and is often about the size of a city block). The average population for a census block in the area represented by the SCCS is approximately 80 people, while the five counties have an average population of 3.765 million.

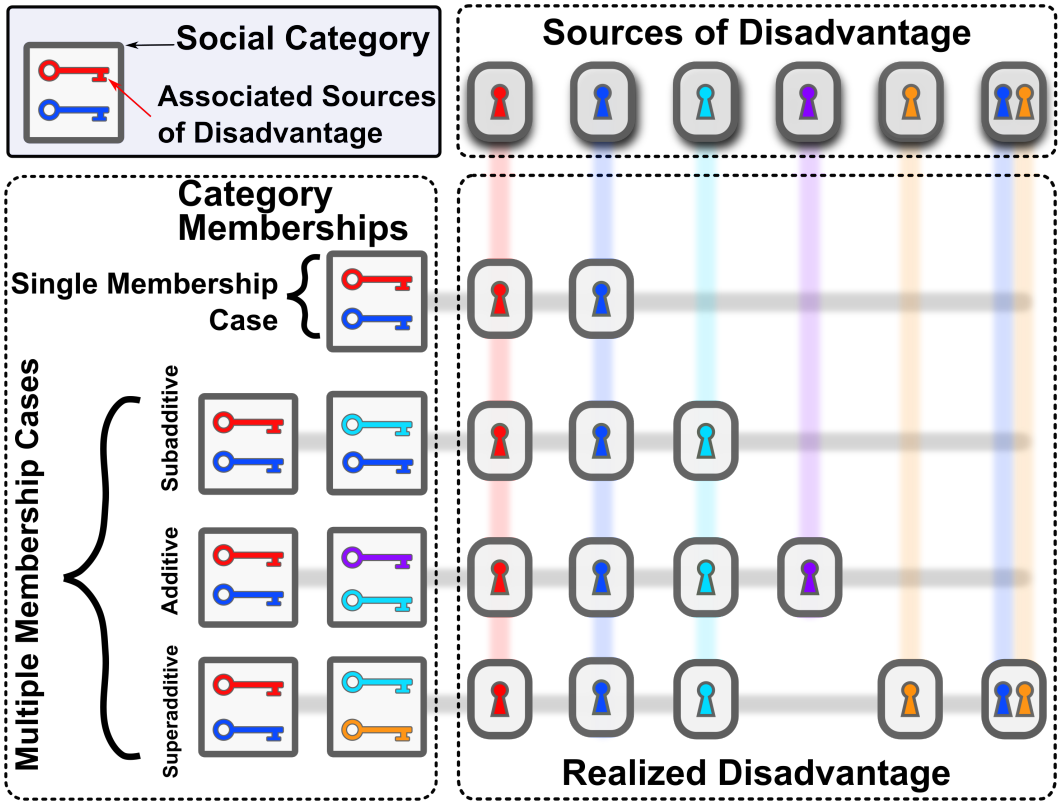


Figure 2: Schematic depiction of the ways in which overlapping social category memberships can lead to different degrees of realized disadvantage. Each social category (i.e. a race/gender/ethnicity category) has a set of associated sources of disadvantage. These sources of disadvantage can combine in a variety of ways. In the subadditive case, overlapping sources of disadvantage only contribute once to the total degree of disadvantage. In the additive case, all sources of disadvantage contribute once to the total amount of realized disadvantage. Under superadditivity, additional disadvantage is "unlocked" due to having multiple sources of disadvantage in distinct social categories.

In order to compute one's exposure to crime (i.e. the response term for this model), we use the number of crime events that occur in a given areal unit, using data from the SCCS data source. From this number of crime events, we consider a group's exposure to crime in that areal unit as

$C_{ijk} = Ep_{ijk}$, where E is the number of violent crime events that occur for that areal unit, and p_{ijk} is the proportion of the total number of people in a three-way race, ethnicity, gender category (for the entire sample) that are present in the areal unit. These values are summed across all areal units to produce an exposure for each of the groups. This exposure functionally behaves as a weighted average of crime exposures in each areal unit. We then predict this exposure using dummy variables for each race, ethnicity and gender category, as well as all two and three way interaction effects using these terms. The White, Non-Hispanic Male category is used as a reference group for these regressions.

Given that these fully specified models are not common in the literature, we present three hypotheses about the nature of the relationship between the explanatory factors and one's exposure to crime, motivated by more general notions of cumulative, intersectional, and saturated mechanisms of disadvantage. These hypotheses are schematically represented by Figure 2. We consider individuals as belonging to one or more *social categories,* reflecting e.g., race, gender, etc. Members of a given social category may be, on average, particularly likely to be exposed to specific *underlying sources of disadvantage*; some such sources may be unique to specific categories, while others may be shared by members of multiple categories. Schematically, Figure 2 depicts social categories as boxes, each of which contains a set of "keys" that "unlocks" particular sources of disadvantage (here, indicated by color). An individual belonging to only one social category receives the keys - and hence the sources of disadvantage - for that category. Where an individual belongs to multiple categories, they inherit the keys from each category they belong. The consequences of this can vary, leading to several hypothetical scenarios.

Our first scenario, represented by the *subadditive* row of Figure 2 involves the case where the sources of disadvantage for an individual's social categories overlap. In this case, having multiple memberships in disadvantaged groups provides more disadvantage than being a member of a single category, but not as much as the independent combination of both groups. Here, the sources of disadvantage *saturate*, and their effect on the individual is subadditive.

Our second scenario, represented by the *additive* row, occurs when there is no overlap in the sources of disadvantage for the categories to which an individual belongs. Here, the total disadvantage is simply the sum of the disadvantage for each category.

Finally, in our third scenario (the *superadditive* row) we consider the possibility that there are sources of disadvantage that require "keys" from multiple categories to unlock. In this case, the total

disadvantage for multiple group memberships can exceed the sum of the group disadvantages, since a joint member is impacted by both the union of the two group sources and additional sources of disadvantage that arise from co-membership. This is often discussed in the literature within the context of *intersectionality* (Crenshaw, 1990), with the notion that belonging to multiple disadvantaged groups can have a substantially greater impact than the independent effects of each membership alone.

In the context of exposure to crime, it is plausible that sources of disadvantage associated with gender, race, and ethnicity could correspond to any of these three scenarios. To quantify this, we specify an *additivity index*, which we use to categorize the relationship for each of our three-way categories included in the model. This index can be defined by:

$A = \frac{\beta_{ijk}}{a+\beta_i+\beta_j+\beta_k+\beta_{ik}+\beta_{jk}+\beta_{ij}}$

where a is the intercept term, and the β terms are the regression coefficients for the one, two and three-way effects. If the three-way effect in this term is zero, then the index will also be zero, which implies a purely additive relationship. Likewise, if the sign of the total two-way effects and three-way interaction term are the same, then the index will be positive, indicating a superadditive relationship. Finally, if the numerator and denominator are of different signs, this index would be negative, which indicates a subadditive relationship. In the rare case where the total two-way effect (denominator) is zero, we define A to be zero.

The magnitude of the index is also informative. Usually, we would expect A to be between -1 and 1, which indicates that the three-way effect is smaller in magnitude to the rest of the effects. However, in the event that A is greater in magnitude than -1 or 1, this indicates that the three-way effect outstrips the combined two-way effects, and would be able to flip the sign of the total effect.

5 Tract Imputation Results

Next, we describe the results of the imputations discussed above, evaluating the overall quality of the imputed arrays. Imputing all 73,057 tracts took 7 hours and 33 minutes and 49 seconds on 30 cores of an Intel Xeon E5-2599 V4 CPU. As the tract-level three-way arrays in the United States are known, we can directly compare the imputed three-way arrays with to the observed data. We use the error metric specified in Section 4.3.1.

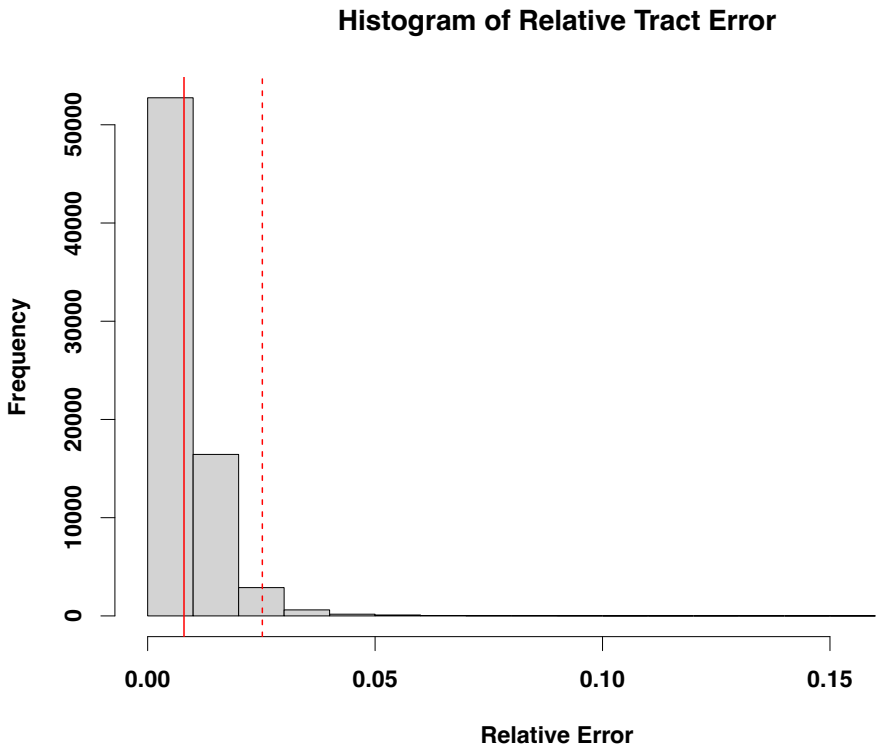


Figure 3: A histogram of relative errors. The solid red line is the mean (0.8%), while the dashed red line is the 97.5^{th} percentile (2.52%)

Array approximation results. This error metric provides support for the quality of the data imputation. Figure 3 describes the distribution of relative errors, showing that most tracts have a very low error rate. Given that the mean relative error is 0.8% and the 97.5th percentile of the error is 2.5%, this imputation schema produces three-way arrays that are excellent proxies for the observed data (with error rates at or below error rates in the Census itself (Khubba et al., 2022)).

We compare these error rates to the rates obtained by the other procedures described in section 4.3.1. We find that the case where we draw directly from the target distribution (using Bunea and Besag's algorithm without simulated annealing) produces very slightly elevated error rates to the ones produced by our updated algorithm. The mean error rate for all tracts in the U.S. is 0.009 (0.9%), while the 97.5th percentile of the error is 0.0283 (2.83%). The arrays produced by this algorithm are produced by the same process that we use for multiple imputation, which we will also show produces similar qualitative results to the single imputation case when doing downstream data

analysis. We thus conclude that there is some gain from annealing to find the mode of the target distribution (versus using an arbitrary draw), but error rates are not very sensitive to this aspect of the algorithm.

Next, when using the expected counts produced by the loglinear models (for our target distribution) as the imputed arrays, we see noticeably elevated error rates. The mean error rate is 0.0124 (1.24%), while the 97.5th percentile of this error distribution is 0.0396 (3.96%). While these error rates are still relatively low, they are roughly 50% higher than the annealed imputation method, and the estimates do not satisfy integer constraints (making them unsuitable for some applications).

For the third case, where we simply construct a valid three-way array that conforms to the integer and marginal constraints, we would expect the error rates to be significantly higher than for the case where we use simulated annealing to produce the most likely three-way array under the target distribution. Indeed, the mean error rate produced by this imputation is 0.047 (4.7%), while the 97.5th percentile of the error is 0.158 (15.8%). This case provides an interesting point of comparison, as it shows that the space of three-way arrays is significantly constrained by the two-way marginals, but despite this, there are still significant improvements that are made through the optimization components of the algorithm. To further visualize the differences between the imputations described here, we plot the error histograms in full here (see Fig 5).

Overall, these results suggest that while constraints are powerful, incorporating distributional target information is still important for getting high-quality approximations. Given that this is done, optimization to ensure that a mode is selected (versus a random draw from the target) is helpful, but less vital. This also implies that our approach is not extremely sensitive to annealing performance, which may be useful in settings for which the cost of high-quality annealing runs is a concern.

Results for downstream analysis. In addition to direct approximation error, we also use the case study described in Section 4.3.2 to evaluate data quality. Our case study examines the effects of disadvantaged social category membership on exposure to crime. As we are using three-way arrays to examine this relationship, we are particularly interested in the three-way coefficients from the regression specified above. The coefficients for both the observed data model and the imputed data model are visualized in Figure 5. For the observed data model, the means and variances are directly computed from 1000 bootstrapped samples of areal units. We use a standard bootstrap sampling

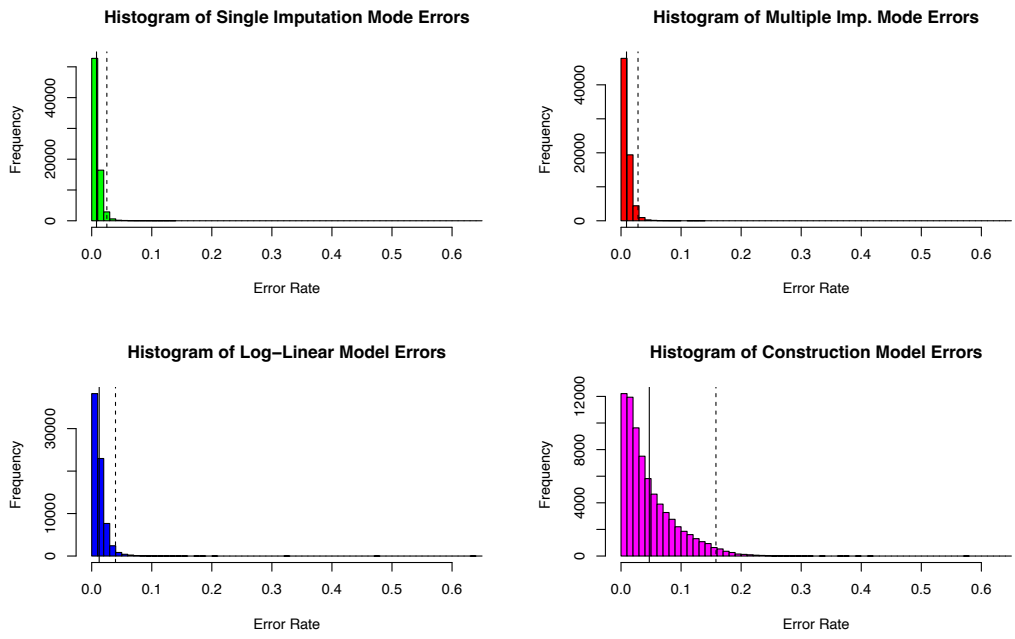


Figure 4: A set of histograms, each describing the error rates for one of the imputations described in Section 5. Solid lines denote the mean error rates, while the dashed lines denote the 97.5th percentile of each distribution. The single and multiple imputation modes for the algorithm we describe both produce high quality data, although the single imputation mode does offer slightly lower error rates. Using expected values from the log-linear model produces low error rates, although these data are not guarunteed to be integer values. The construction model produces a wide range of errors at significantly elevated rates compared to the full imputation technique we describe in the paper.

design for the observed data model. For the imputed data model, we examine the effect of using the algorithm in single imputation mode vs. multiple imputation mode. In single imputation mode, we draw a single array for each areal unit. Then, we sample areal units using the standard bootstrap design. In the multiple imputation mode, we use a slightly different sampling method. For each of the 2000 bootstrapped samples, we draw a set of arrays from the distribution defined by our target distribution. The Markov chains used to draw from this distribution were seeded with a draw from a single imputation run of the algorithm to ensure that the chains were burned in adequately. Then, we draw n arrays from that areal unit, where n is the number of times that areal unit has been drawn for that bootstrap sample. We then use the quantile method to compute the distribution of each coefficient.

The three-way coefficients from Figure 5 almost all overlap with zero, indicating mostly *additive effects*. In addition, the simulated and observed distributions of three-way coefficients all have

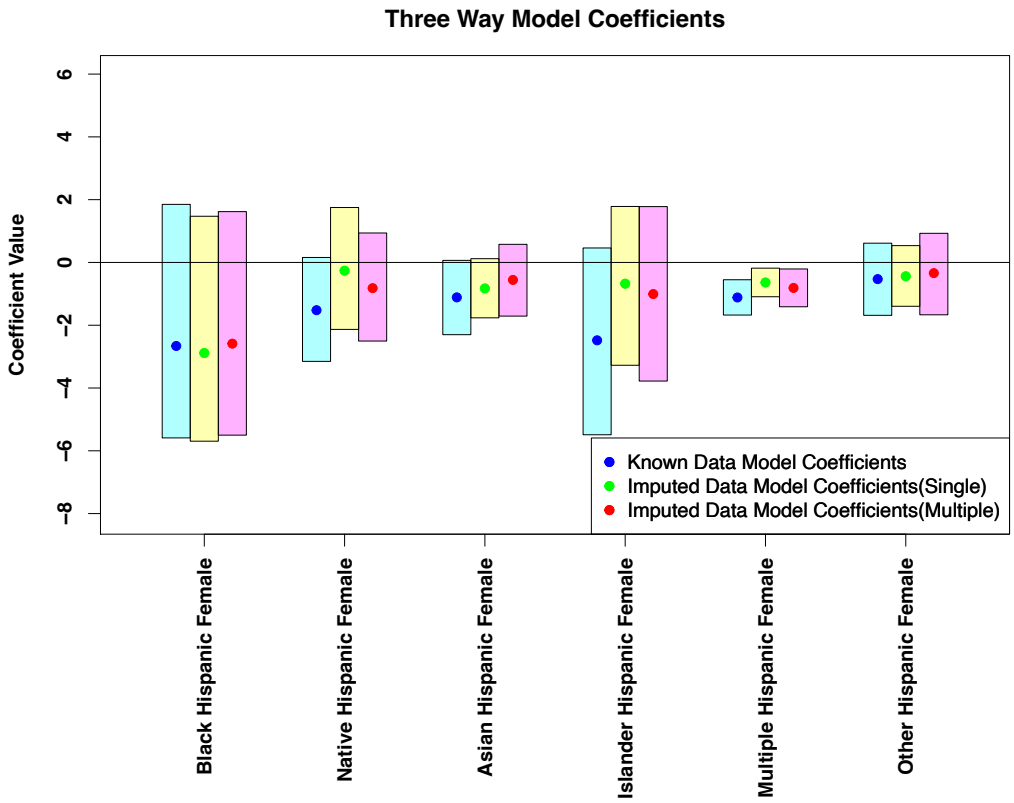


Figure 5: A plot of three-way effects where the blue points are the coefficients of the known model with cyan 95% simulation intervals, and the red points are coefficients of the multiply imputed model with magenta 95% simulation intervals. The green points and yellow 95% simulation intervals are for a model that uses a bootstrap design, but with single imputation rather than multiple imputation. Known data simulation intervals were computed with 2000 bootstrap iterations using the quantile method. Imputed model intervals (red) were computed using a set of MCMC samples that utilize the multiple imputation mode of the algorithm, while the yellow intervals use a single imputation mode of the algorithm.

significant overlap with each other. Further, in interpreting the effects, a researcher would obtain similar qualitative results from interpreting the observed and simulated models (i.e. the effects significantly different from zero are the same). This indicates that qualitatively, the simulated arrays are similar in nature to the observed arrays, and would not introduce significant error into downstream analysis. Additionally, while the single imputation produces coefficients that tend to be slightly closer to zero, both single and multiple imputation modes produce similar results to the observed data model.

We also examine the patterns in the coefficients reported in Figure 5, with respect to their additivity indices. For each of the coefficients reported from the analysis done with the bootstrapped

samples from the observed three-way arrays, we compute the additivity index from Section 4.3.2. The distribution of these additivity indices is reported in Figure 6. As expected from a cursory examination of the coefficients in Figure 5, the majority of the coefficients exhibit an additive pattern. For the Black Hispanic Female and Pacific Islander Hispanic Female coefficients, there are some cases of subadditivity present, but for the most part the three-way coefficients describe an additive relationship between disadvantaged category membership and exposure to crime at the tract level. Notably, we find no sign of systematic superadditivity at this level of aggregation.

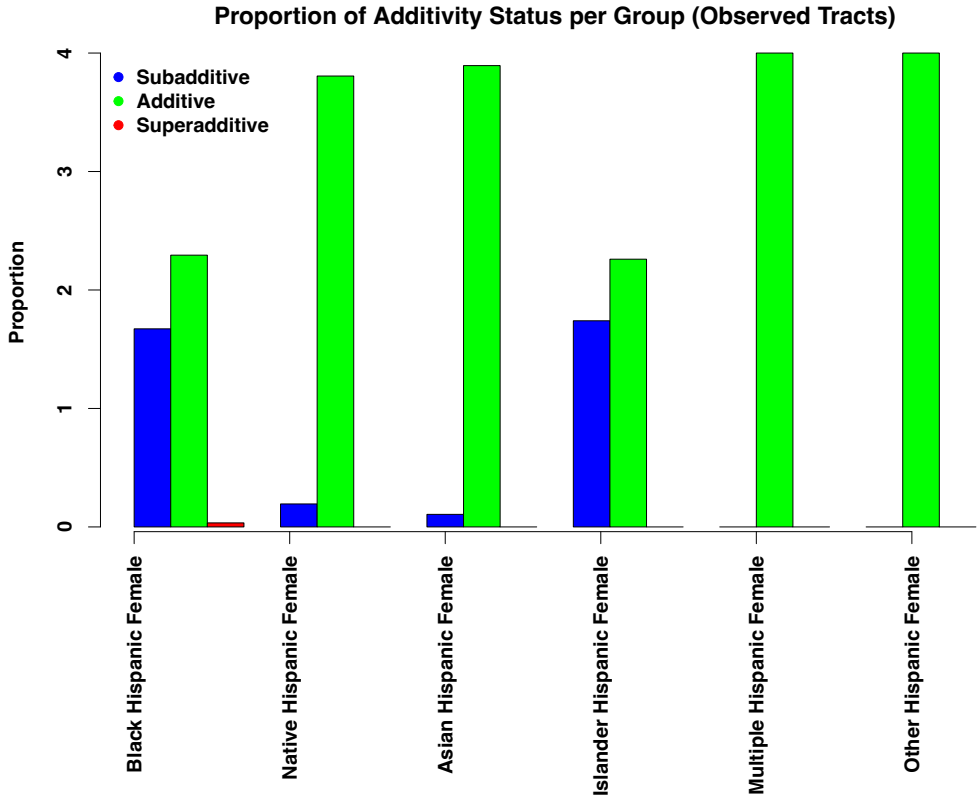


Figure 6: Additivity indices for each of the three-way categories in the model across 2000 bootstrap iterations. Additive values were ones in which the three-way effect size was less than 5% of the combined one and two-way effects.

5.1 Block Level Imputation Results

While the tract-level analysis suggests additive effects to be predominant, it is possible that this is an artifact of aggregating over locally heterogeneous units. Although the full data needed to replicate the observed-data analysis at the block level is not available, we can do so using our imputation

scheme. For a single imputation, we impute the 710,145 census blocks in California on 30 cores of an Intel Xeon E5-2599 V4 CPU in 12 hours, 53 minutes, and 42 seconds. We hypothesize that the higher areal unit/second imputation rate is likely due to the lower population of census blocks in comparison with census tracts. There are also more census blocks that can be trivially solved (see Section 4.2) than census tracts. The full three-way arrays at the census block level are not available, so we are unable to compute the same error metrics that we use for the census block imputation.

Given the imputed block-level arrays, we once again examine the relationship between gender, ethnicity, and race on exposure to crime. We apply the same basic procedure as in the tract level case study. However, rather than measuring crime in the specific block (which is too small a unit of analysis to provide a reasonable notion of exposure), we measure crime in a 1km buffer around each focal block. Once again, we use multiple imputation to get a set of potential arrays for each areal unit, drawing from the target distribution specified at the tract level.

The three-way effects from the block level exposure to crime are summarized in figure 7. We used 500 bootstrapped samples from the crime and areal unit data to compute the simulation intervals for this plot. The patterns in these coefficients generally match the patterns from the tract level analysis, but the magnitude of these coefficients is much greater. None of the three-way coefficients intervals contain zero, representing significant three-way effects. We use the same multiple imputation sampling that is described above to generate these estimates.

As as the observed effects are consistently negative, we expect that there is less of a strongly additive pattern at the block level. Figure 8 depicts the patterns in the additivity index for the three-way coefficients. Almost all of the coefficients exhibit a strongly subadditive pattern, with only the Other race Hispanic Female and Asian Hispanic Female categories having some additive indices. We thus find that, at fine spatial scales, the relationship between exposure to crime and disadvantaged group memberships is *subadditive*, implying that the sources of disadvantage that are associated with group memberships overlap. This runs counter to the common intuition that disadvantage is compounded across social categories, but is mechanistically sensible: while many things can lead to e.g. living in poor housing, or having a large number of potential offenders nearby, once one acquires such a source of disadvantage there is a limit to how much additional impact it can have. Thus, the sources eventually saturate, with diminishing marginal effects. This nonlinear effect is lost when data is aggregated to the tract level, as would be necessary without the ability to impute at the block level.

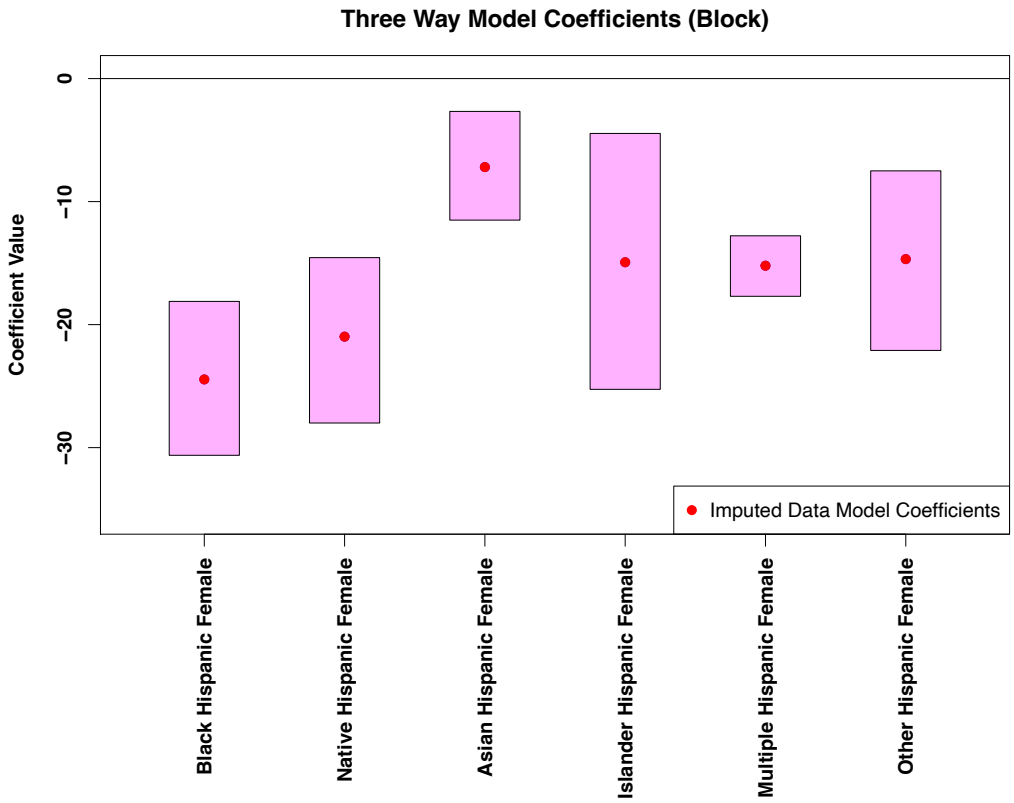


Figure 7: Estimates of three-way coefficients at the census block level. The red points are the mean estimates across 500 bootstrap intervals, with the magenta region representing the 95% simulation intervals.

While the pattern of additivity indices is different between the block and tract level analyses, we note that the general pattern of the coefficients matches. We do not therefore see radical differences across scales, but rather subtle variations that can be obscured by averaging. As noted, however, those variations can lead to distinct substantive conclusions about that nature of disadvantage in crime exposure.

6 Discussion

We have implemented and tested an imputation framework for nested areal units, showing that it produces high quality data for three-way arrays that contain count data. The algorithm specified in this paper should produce high quality data for arrays in which all entries are non-negative integers. Likewise, we are able to leverage data at a higher level of geographic aggregation to optimize the

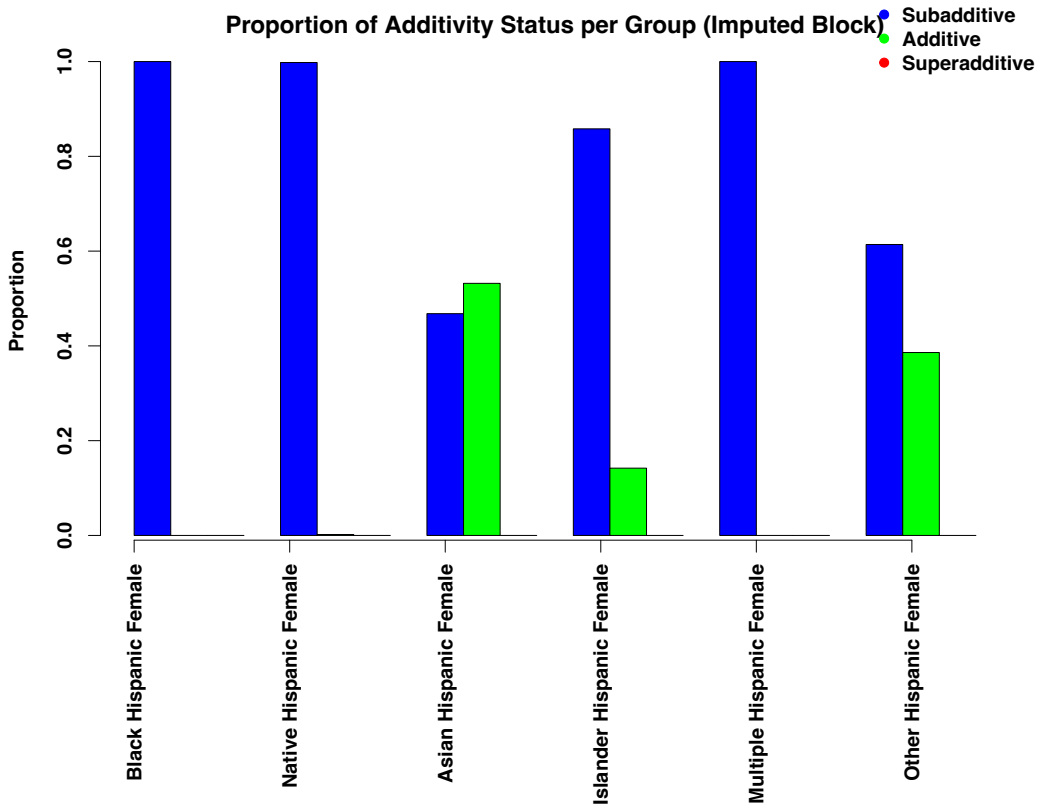


Figure 8: Additivity Indices for the three-way block coefficients. Coefficients are in the additive category when the three-way coefficient is less than 5% of the other combined effects.

configuration of an imputed array to what we expect the correlations between array cells to be.

With the recent push in many fields within the social sciences for measures constructed at smaller geographic scales, along with the limited availability of some data at such small geographic scales, our imputation algorithm may be applicable in a range of settings. For example, while our case study showed a generally similar pattern in crime exposure for residents of different demographic groups whether measured in census tracts or the smaller spatial unit of blocks, we nonetheless saw sharper and stronger patterns when using the smaller geographic units. Given the spatial segregation of residents across the landscape at varying spatial scales, measuring such effects at smaller geographic scales is arguably substantively important for addressing such research questions. The spatial averaging that occurs when aggregating to larger geographic units has the risk of obscuring such patterns that we were able to observe after imputing the data to blocks.

While the implementation of this algorithm represents a substantial step forward in three-way array imputation, especially with the constraints we describe above, the problem of imputing high

order arrays is still difficult. For example, while Bunea and Besag (2000) claim that for the three-way crosstab imputation problems without two-way marginal constraints (i.e. with independent one-way margins), the Monte Carlo method is amenable, this remains a scenario where the formulae are not yet derived and the algorithms are not yet implemented. They also point out that their algorithm is specific to the three-way array case, and while the basic move could, in theory, be adapted to a higher order array, the problem cannot be solved in a "plug and play" fashion. A new transition set must be derived for higher-order arrays, as the transition rule used here is specific to the three-way case (and, in general, each order requires a new set of basic moves, associated with its respective symmetry group). Integrating higher order marginal constraints into models for simulating high-order array data would also be a valuable next step in this line of research.

Another open question in this area is the provable irreducibility of the underlying Markov chain used for the multiple imputation case; while the construction algorithm provided here is verifiable, and the optimization and imputation algorithms guarantee that the result is margin-preserving, the basic move of Bunea and Besag has not been proved to be irreducible in all cases. Their original paper proves irreducibility for any $I \times J \times K$ array such that one of the dimensions has cardinality 2. Irreducibility is thus ensured for the cases studied here, or any other population data using e.g. a two-class sex tabulation. Subsequent work by Lee (2018) provides a proof for the $3 \times 3 \times K$ case, as well as simulation studies suggesting that the property is preserved for cardinalities $4 \times 4 \times 4$ and higher. It thus appears likely that the property holds for all three-way arrays, though this is still unproven. Failure of irreducibility would imply that the Markov chain would not explore the entire state space of possible arrays, thus possibly (1) finding a point estimate that suboptimally captures three-way correlations, or (2) in the multiple-imputation case, providing an imperfect approximation to the target array distribution. (It would not, however, lead to invalid imputations.) Care should thus be exercised when using this method for multiple imputation on arrays that violate the cardinality conditions, when a high level of precision is required.

We observe that the model that we use in this paper to simulate three-way array data is highly scalable. For small areal unit estimation, our imputation scheme does not require information on adjacent units imputed values, so each areal unit can be estimated independently. This has substantial computational benefits, as large sets of areal units can be imputed in parallel, which provides a significant decrease to overall runtime. We were able to simulate three-way distributions for all tracts in the United States, as well as all census blocks in California in less than 24 hours,

showing that the algorithm can be employed in over very large regions. As both the construction and optimization problems contribute to runtime, it is likely that arrays with lower population would increase imputation speed, as the construction algorithm will converge more quickly with fewer people.

We note that release of small areal unit data often reflects a "tug-of-war" between advocates of openness, transparency, and data quality (on the one hand) and privacy (on the other). Each faction cites a range of arguments in its favor (often with a certain degree of zealotry), and we here limit ourselves to commenting on implications for the imputation problem. Three-way imputation applied to valid data may or may not allow data identification at a given level of confidence, depending on cell counts; for the scenario studied here, high-confidence identification must come from the two-way marginals, as the higher-order correlation structure is both estimated and approximate. Techniques such as differential privacy can be employed by data collectors to design perturbed marginals that provide guaranteed bounds on identifiability, and algorithms such as those shown here may be useful for verifying the results of such constructions (and ensuring that they still lead to valid arrays). Similar validation applications are possible for privacy-preserving techniques based on e.g. areal unit aggregation (where units are merged until they no longer permit identification beyond the specified level of confidence). Given that data-perturbing methods like differential privacy pose significant data quality concerns, another use for imputation methods of the type shown here is to ensure that the perturbed data yields imputed arrays that are still appropriate for downstream analysis. Particularly given the importance of neighborhoods, blocks, and other small units for social processes related to social disadvantage, we observe that obfuscation methods that induce systematic bias in small scale structure have the potential to negatively affect policy relevant research impacting vulnerable communities. It is hoped that obfuscators will leverage imputation and related methods to help verify that their modifications will not have such downstream effects.

The imputation techniques introduced here could also be extended in several ways. First, we consider the case where there are additional spatial dependencies amongst the areal units. For this case, the algorithm could be extended by generating a target distribution from both the areal unit immediately higher in the spatial hierarchy from the target unit, but also from that unit's neighbors. A natural approach is to estimate τ^H_{ijk} using a spatial smoother at the level of higher-order units, then employing this (rather than the τ^H_{ijk} based only on the parent unit) for lower-level imputation. Directly incorporating autocorrelation at the lower level is also possible, but would require a more

complex, multi-level and would be less amenable to parallelization. Both are potentially fruitful directions for further work.

Likewise, this technique could also be extended to the case in which the areal units are not perfectly hierarchical. The most obvious direction to extend the algorithm would be similar to the extension described for using multiple parent geographies. In this case, it may make sense to average the correlation structure of parent geographies that overlap with the target areal units. A more radical proposal of this type is suggested for cases in which complete three-way information is available for some units, while only two-way information is available for others. In this case, an interesting option is to train a kernel learner (Scholkopf and Smola, 2001) or similar predictive algorithm to predict the lower-level τ_{ijk} coefficients from observed marginals and other spatial and contextual data; the trained algorithm can then be employed to predict $\hat{\tau}_{ijk}$ directly, as opposed to using τ^H_{ijk} as a proxy. Although kernel learning suggests itself due to its interpretation in terms of a similarity function, other methods could be used as well.

On a final, substantive note, we observe that our sample application to exposure-to-crime data suggests that disadvantage in this context is largely subadditive: notably, we do not see the super-additive effects often presumed (but less often tested) in sociological discussions of intersectionality. We also observe that this subadditivity is largely masked at higher levels of geography (though we do not see evidence of superadditive effects there, either). While it is possible that this is peculiar to the case of crime exposure, the mechanistic interpretation discussed here would suggest that the phenomenon may be much more common. A more systematic investigation of when and how often disadvantage is additive, subadditive, and superadditive across different contexts and for different types of disadvantage would greatly illuminate theory in this area, and may also inform policy interventions. Regardless, our findings reinforce the value of fine-grained spatial data for accurate assessment of local social processes.

7 Conclusion

We here specified and demonstrated an algorithm for imputing three-way array data within a hierarchically nested context. This imputation problem is challenging, as it is constrained by the two-way marginal structure of the array, an integer constraint, as well as needing to be optimized with respect to higher order array data. We provide a scalable, robust technique to impute these

three-way arrays that relies on Markov Chain Monte Carlo and simulated annealing strategies.

In a test imputation of all tracts in the United States, simulated data from our algorithm produced remarkably low error rates. At the tract level, we observed a mean allocation error of approximately 0.8%, with nearly all tracts having errors below 2.5%. Such errors are better than or comparable to error levels in the Census itself (Khubba et al., 2022), suggesting that imputation is unlikely to be a dominant source of error in subsequent analyses. Likewise, in a case study that examines three-way categories exposure to crime, we found that both imputed data and observed data produced similar conclusions about the relationship between disadvantaged category membership and exposure to crime. Combined, both of these metrics show that the imputed arrays are very similar to observed arrays, and can be used in downstream analyses without introducing significant error.

As sketched above, there is considerable room for further work on the imputation of higher-order array data embedded in spatial hierarchies. With the proliferation of these nested data structures, methods that allow for the use of data at low levels of geographic aggregation where data may be incomplete are particularly valuable.

References

Alba, Richard D, John R Logan, and Paul E Bellair. 1994. "Living with crime: The implications of racial/ethnic differences in suburban location." *Social Forces* 73:395–434.

Bennett, Richard J, Robert P Haining, and Daniel A Griffith. 1984. "The problem of missing data on spatial surfaces." *Annals of the Association of American Geographers* 74:138–156.

Boessen, Adam and John R Hipp. 2015. "Close-ups and the scale of ecology: Land uses and the geography of social context and crime." *Criminology* 53:399:426.

Bunea, Florentina and Julian Besag. 2000. "MCMC in IxJxK Contingency Tables." *Monte Carlo Methods* 26:25.

Clogg, Clifford C and Scott R Eliason. 1987. "Some common problems in log-linear analysis." *Sociological Methods & Research* 16:8–44.

Cohen, Michael Lee and Xiao Di Zhang. 1988. "The difficulty of improving statistical synthetic estimation." .

Crenshaw, Kimberle. 1990. "Mapping the margins: Intersectionality, identity politics, and violence against women of color." *Stan. L. Rev.* 43:1241.

Darroch, John N and Douglas Ratcliff. 1972. "Generalized iterative scaling for log-linear models." *The annals of mathematical statistics* pp. 1470–1480.

Espuny-Pujol, Ferran, Karyn Morrissey, and Paul Williamson. 2018. "A global optimisation approach to range-restricted survey calibration." *Statistics and Computing* 28:427–439.

Facebook Connectivity Lab, Center for International Earth Science Information Network, CIESIN, and Columbia University. 2016. "High Resolution Settlement Layer (HRSL)." Source imagery for HRSL © 2016 DigitalGlobe, accessed May 31 2022.

Graham, Patrick, Jim Young, and Richard Penny. 2009. "Multiply imputed synthetic data: Evaluation of hierarchical Bayesian imputation models." *Journal of Official Statistics* 25:245.

Jaynes, Edwin T. 1982. "On the rationale of maximum-entropy methods." *Proceedings of the IEEE* 70:939–952.

Khubba, Shadie, Krista Heim, and Jinhee Hong. 2022. "National Census Coverage Estimates for People in the United States by Demographic Characteristics: 2020 Post-Enumeration Survey Estimation Report." US Census Report PES20-G-01.

King, Gary, Ori Rosen, and Martin A Tanner. 1999. "Binomial-beta hierarchical models for ecological inference." *Sociological Methods & Research* 28:61–90.

Kubrin, Charis E and John R Hipp. 2016. "Do fringe banks create fringe neighborhoods? Examining the spatial relationship between fringe banking and neighborhood crime rates." *Justice Quarterly* 33:755–784.

Kubrin, Charis E, John R Hipp, and Young-An Kim. 2018. "Different than the sum of its parts: Examining the unique impacts of immigrant groups on neighborhood crime rates." *Journal of Quantitative Criminology* 34:1–36.

Lee, Seungchan. 2018. "Markov Chain Monte Carlo and Exact Conditional Tests with Three-Way Contingency Tables." Technical report, Naval Postgraduate School.

Logan, John R and Brian J Stults. 1999. "Racial differences in exposure to crime: The city and suburbs of Cleveland in 1990." *Criminology* 37:251–276.

Luna, Angela, Li-Chun Zhang, Alison Whitworth, and Kirsten Piller. 2015. "Small area estimates of the population distribution by ethnic group in England: a proposal using structure preserving estimators." *Statistics in Transition new series* 16:585–602.

McHugh, Mary L. 2013. "The chi-square test of independence." *Biochemia medica: Biochemia medica* 23:143–149.

McNulty, Thomas L. 1999. "The residential process and the ecological concentration of race, poverty and violent crime in New York City." *Sociological Focus* 32:25–42.

Molina, Isabel and JNK2010 Rao. 2010. "Small area estimation of poverty indicators." *Canadian Journal of Statistics* 38:369–385.

Mooney, Stephen J, Michael DM Bader, Gina S Lovasi, Kathryn M Neckerman, Andrew G Rundle, and Julien O Teitler. 2020. "Using universal kriging to improve neighborhood physical disorder measurement." *Sociological Methods & Research* 49:1163–1185.

Moretti, Angelo and Adam Whitworth. 2020. "Development and evaluation of an optimal composite estimator in spatial microsimulation small area estimation." *Geographical Analysis* 52:351–370.

Morrison, Peter A. 1971. "Demographic Information for Cities: A Manual for Estimating and Projecting Local Population Characteristics." .

Pfeffermann, Danny and Solange Correa. 2012. "Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation." *Biometrika* 99:457–472.

Pfeffermann, Danny et al. 2013. "New important developments in small area estimation." *Statistical Science* 28:40–68.

Rose, A., J. McKee, K. Sims, E. Bright, A. Reith, , and M. Urban. 2021. "LandScan Global 2020 (Data set)."

Rosen, Ori, Wenxin Jiang, Gary King, and Martin A Tanner. 2001. "Bayesian and frequentist inference for ecological inference: The R× C case." *Statistica Neerlandica* 55:134–156.

Rubin, Donald B. 1996. "Multiple imputation after 18+ years." *Journal of the American statistical Association* 91:473–489.

Scholkopf, Bernhard and Alexander J. Smola. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press.

Steinberg, Joseph. 1979. "Synthetic Estimates for Small Areas: Statistical Workshop Papers." *National Institute on Drug Abuse Research Monograph Series* 24:282.

Thomas, Loring J, Peng Huang, Fan Yin, Xiaoshuang Iris Luo, Zack W Almquist, John R Hipp, and Carter T Butts. 2020. "Spatial heterogeneity can lead to substantial local variations in COVID-19 timing and severity." *Proceedings of the National Academy of Sciences* 117:24180–24187.

Thomas, Loring J., Peng Huang, Fan Yin, Junlan Xu, Zack W. Almquist, John R. Hipp, and Carter T. Butts. 2022. "Geographical Patterns of Social Cohesion Drive Disparities in Early COVID Infection Hazard." *Proceedings of the National Academy of Sciences* 119:e2121675119.

Tillman, Balint, Athina Markopoulou, Carter T. Butts, and Minas Gjoka. 2019. "2K+ Graph Construction Framework: Targeting Joint Degree Matrix and Beyond." *IEEE/ACM Transactions on Networking* 27:591–606.

Vermunt, Jeroen K, Joost R Van Ginkel, L Andries Van der Ark, and Klaas Sijtsma. 2008. "9. Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis." *Sociological Methodology* 38:369–397.