Structure, Dynamics, and Inference in Networks

by

Philip S. Chodrow

Submitted to the Operations Research Center
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

© Massachusetts Institute of Technology 2020. All rights reserved.

Author .....
Operations Research Center
May, 2020

Certified by .....
Patrick Jaillet
Dugald C. Jackson Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by .....
Marta González
Associate Professor of Urban Planning, UC Berkeley
Thesis Supervisor

Accepted by .....
Georgia Perakis
William F. Pounds Professor of Management
Co-Director, Operations Research Center

Structure, Dynamics, and Inference in Networks

by

Philip S. Chodrow

Submitted to the Operations Research Center
on May, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

Networks offer a unified, conceptual formalism for reasoning about complex, relational systems. While pioneering work in network science focused primarily on the ability of "universal" models to explain the features of observed systems, contemporary research increasingly focuses on challenges and opportunities for data analysis in complex systems. In this thesis we study four problems, each of which is informed by the need for theory-informed modeling in network data science.

The first chapter is a study of binary-state adaptive voter models (AVMs). AVMs model the emergence of global opinion-based network polarization from localized decision-making, doing so through a simple coupling of node and edge states. This coupling yields rich behavior, including phase transitions and low-dimensional quasistable manifolds. However, the coupling also makes these models extremely difficult to analyze. Exploiting a novel asymmetry in the local dynamics, we provide low-dimensional approximations of unprecedented accuracy for one AVM variant, and of competitive accuracy for another.

In the second chapter, we continue our focus on fragmentation in social systems with a study of spatial segregation. While the question of how to measure and quantify segregation has received extensive treatment in the sociological literature, this treatment tends to be mathematically disjoint. This results in scholars often re-proving the same results for special cases of measures, and grappling with incomparable methods for incorporating the role of space in their analyses. We provide contributions to address each of these issues. With respect to the first, we unify a large body of extant segregation measures through the calculus of Bregman divergences, showing that the most popular measures are instantiations of generalized mutual informations. We then formulate a microscopic measure of spatial structure – the local information density – and prove a novel information-geometric result in order to measure it on real data in the common case in which the data is embedded in planar network. Using these tools, we are then able to formulate and evaluate several network-based regionalization algorithms for multiscale spatial analysis.

We then take up two questions in null random graph modeling. The first of these develops a family of null random models for hypergraphs, the natural mathemati-

cal representation of *polyadic* networks in which multiple entities interact simultaneously. We formulate two distributions over spaces of hypergraphs subject to fixed node degree and edge dimension sequences, and provide Markov Chain Monte Carlo algorithms for sampling from them. We then conduct a sequence of experiments to highlight the role of hypergraph configuration models in the data science of polyadic networks. We show that (a) the use of hypergraph nulls can lead to directionally different hypothesis-testing than the use of traditional nulls and that (b) polyadic nulls support richer and more complex measurements of graph structure. We close with a formulation of a novel measure of correlation in hypergraphs, as well as an asymptotic formula for estimating its expectations under one of our configuration models.

In the final chapter, we study the expected adjacency matrix of a uniformly random multigraph with a fixed degree sequence. This matrix is an input into several common network analyses, including community-detection and mean-field theories of spreading properties on contact networks. The actual structure of this matrix, however, is not well understood. The main issues are (a) the combinatorial complexity of the space on which this random graph is defined and (b) an erroneous folk-theorem among network scientists which stems from confusion with related models. By studying the dynamics of a Markov chain sampler, we prove a sequence of approximations that allow us to estimate the expected adjacency matrix – and other elementwise moments – using a fast numerical scheme with qualified uniqueness guarantees. We illustrate using a series of experiments on primary and secondary school contact networks, showing order-of-magnitude improvements over extant methods.

We conclude with a description of several directions of future work.

Thesis Supervisor: Patrick Jaillet
Title: Dugald C. Jackson Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Marta González
Title: Associate Professor of Urban Planning, UC Berkeley

Acknowledgments

There are many people and organizations who made possible my graduate career at MIT, including (but certainly not limited to) the work contained in this thesis.

Marta González and Patrick Jaillet both contributed helpful ideas and advice, within and beyond research. My success thus far is in large part through their guidance, and I am grateful. Thank you Marta, and thank you Patrick. Peter Mucha was generous with his time and mentorship far beyond any reasonable duty. I suspect that my opportunity for the next stage of my career is in no small part thanks to his advocacy. Thank you Peter. David Gamarnik made time to read my writing and support my career advancement in the midst of a very busy and momentous time. Thank you David.

Though my work fell somewhat outside the standard fare at the MIT Operations Research Center, the ORC nevertheless offered an extremely stimulating environment in which to learn, think, teach, and work. I am grateful to the codirectors, students, and faculty who make it so. I would also like to thank the students at the Northeastern Network Science Center, who welcomed me in and offered a second, welcoming intellectual community with interests closely aligned with my own.

One can't forget about the money. Marta and Patrick both offered me financial support that enabled me to work independently and focus on problems that interested me. I would especially like to note Marta's generosity in sending me to conferences and other meetings. Chapters 3 to 5 all sprang from my participation in scholarly gatherings, as did many friendships and ongoing collaborations. I also received support in the final three years of my PhD through a Graduate Research Fellowship from National Science Foundation, under award number 1122374.

I am deeply grateful for my friends and loved ones who supported me in this time. I would especially like to thank my wife, Dr. Charlotte Rose Morris-Wright. She knows why.

Finally, I would like to dedicate this thesis to the memory of my father, Dr. Don Chodrow. He, a physicist, would have greeted with both pride and good-natured

irritation the news that his son was pursuing a career in mathematics.

Contents

List of Figures

List of Tables

Chapter 1

Network Theory, Network Data, Network Science

> "Hegel seems to me to be always wanting to say that things which look different are really the same. Whereas my interest is in showing that things which look the same are really different. I was thinking of using as a motto for my book a quotation from King Lear: *I'll teach you differences.*"
>
> Ludwig Wittgenstein [139]

The mathematical theory of graphs can trace its history as far back as Euler, and the mathematical theory of *random* graphs as far back as the seminal work of Erdős and Rényi [69]. The disperse, interdisciplinary movement called "network science" is much more recent. Roughly speaking, *network science* refers to scientific activity animated by the following theses:

(a) Many phenomena in the natural, social, and engineered worlds include structures ("networks") naturally modeled in the language of graphs and their generalizations.

(b) The study of the structure and functions of these networks can both lead to insights within scientific domains and facilitate comparisons between them.

The first papers clearly recognizable as "modern" network science appeared in the late 1990s. It is useful to contrast three that appeared around this time. In 1998, Watts and Strogatz [196] formulated one of the first empirically-motivated random graph models. While previous models such as those of [69] and [37, 137] were generally driven by theoretical concerns, the Watts-Strogatz model was explicitly designed to reproduce two nontrivial properties of real-world networks: small diameters and local clustering. This emphasis on designing models that could account for empirically-observed phenomena is arguably what distinguishes modern network theory from the "pure" forms of graph theory, random graph theory, and statistical physics. A year later, Barabási and Albert [16] published a highly influential study in a similar vein. They observed power-law regularities in the degree-distributions of networks across social, biological, and engineered domains. Like Watts and Strogatz, they then offered a simple mechanistic model that could reproduce stylized versions of these regularities.

Almost contemporaneously, Brin and Page introduced [41] the original PageRank algorithm for ranking results in the Google search engine. PageRank is naturally formulated in terms of a modified random walk on a network substrate. This algorithm has had enormous influence on the development of technology and data analysis in the first two decades of the twentieth century. Perhaps surprisingly, however, PageRank did not play a prominent role in mainstream network science in the first decade of the 21st century. The theory-driven Watts-Strogatz (WS) and Barabási-Albert (BA) papers had much greater direct influence in the broad community of network scientists. In one standard text, published in 2010, PageRank receives roughly four of over 700 pages, while the Watts-Strogatz small-world model occupy over ten pages (553-565) and generative models of power-law degree sequences over fifty (487-541) [150]. This may reflect the backgrounds in mathematics and theoretical physics of many of the pioneers in network science, or the relative scarcity of data and computing power available to scholars at the time.

Twenty years after these developments, network science is undergoing a transition – "a collective soul-searching" in the words of one prominent scholar [102] – away from theory-driven approaches and towards data science. This shift is perhaps most

visible at the flagship conference of the Network Science Society, where the satellite sessions on statistical inference and machine learning in networks each typically draw the largest audiences of any non-plenary conference event. Another anecdotal indication is provided by a conversation at a recent network science workshop, historically collocated with a conference on dynamical systems, as to whether the next workshop should be collocated with a new conference on data science. These transitions are likely driven in part by the increasing availability of data, computational hard, and fast algorithms for extracting insight from networked data sets.

The gradual trend toward data science has several intellectual subcurrents. One of these is increasing skepticism toward the concept of *universality* in complex systems. Universality refers to the insight, first formulated in theoretical statistical physics, that similar macroscopic properties are often observed in systems that differ in their microscopic details. The notion of universality can only be rigorously formulated in abstracted, infinite-size limits, and therefore applied only figuratively to concrete, finite-sized systems. Networks, as interconnected systems composed of many microscopic components, are natural systems in which to seek for universal properties. Indeed, the observations of BA, for example, can be taken to suggest that power-law degree distributions are universal across a broad class of socio-technical networks. As network data science has matured, the relevance of universality in our understanding of empirical systems has however been called into question. Power laws are perhaps the most visible example, with methodological innovations in data analysis [55, 42, 198, 207, 83] driving ongoing controversy concerning their true prevalence in networked systems; [102] provides a useful review. These and other papers suggest, not that universality is flawed as a concept from theoretical statistical physics, but rather that it may be of limited utility when studying empirical networks. One of the great contributions of theoretical statistical physics, dominant in the early days of network science, has been to provide consolidating theories that allow us to see similarities between disparate collective phenomena. The emergence of network data science holds different promise. In contrast to universality, data provides specificity. With data, we have the opportunity to see *differences* between superficially similar

phenomena.

While these developments may be distressing to those seeking universal theories in the study of complex systems science, the prospects for theoretical progress have never been so expansive. A first class of theoretical work is methodological: we must have mathematically and statistically grounded algorithms for extracting insight from complex data sets. A second class is substantive: having understood the particularities of networks in specific scientific domains, we require a theory of their operation conversant with the data. Two trends in the development of data-informed network theory stand out. The first is the increasing use of data additional to the "pure" network structure, though this data is sometimes referred to rather perjoratively as "metadata" or "annotations" [161, 145, 49]. Two classes of metadata have received special attention, both taking the form of labels attached to edges. A *temporal network* [104, 166] is a network with edges endowed with timestamps. Temporal networks are suitable models of systems in which edges represent discrete, time-bounded interactions between agents. Similarly, many data sets contain edges of different qualitative characters. For example, one might wish to study the spread of information over an aggregate network that includes both "in-person" social contacts and online social contacts. A *multiplex* or *multilayer* [34, 117] network consists of a network in which edges are sorted into distinguished sets, or "layers." A huge body of recent literature in network methodology has focused on inference and mining of temporal and multilayer networks. See the citations and citing papers of [104, 117, 141, 166] for helpful reviews. On a somewhat separate strand, recent attention has also been paid to principled methods for incorporating node attributes in standard network data analyses [161, 145].

A second, related development has been increased emphasis on the role of the domain of data in determining the relevant forms of analysis. Early papers such as [16] and [152] often conducted comparative analyses of data across many different domains. While this style of analysis remains relevant, many recent papers have highlighted that the domain-specific generating process of data sets can make certain analyses more or less appropriate. The ongoing development of mesoscale-

structure detection – also called "community-detection" and "graph clustering" in various communities – offers a useful example. There are at least three broad classes of method, each best suited to different application types. Inferential approaches based on stochastic blockmodels [162, 5, 113, 100] are ideal when domain-specific considerations suggest the existence of latent node attributes that govern observed connectivity patterns. Approaches based on the compression of trajectories [62, 178] are appropriate for model-reduction tasks in which the dynamics of the trajectories are physically motivated. Optimization-based approaches such as Laplacian spectral clustering [208, 187] are often based in engineering contexts in which the interpretation of a partition is less important than its performance under a suitably-specified objective function. These three broad approaches can overlap; for example, the much maligned method of modularity-maximization can be derived from any of these three points of view [144, 62].

Each chapter of this thesis is informed by the evolving dialog between network theory and network data science, and the mathematical challenges generated by that dialog. There is an emphasis throughout on the role of domain considerations in modeling decisions and analytical methodology, and the ability of carefully-crafted theory to highlight interesting differences between disparate networks. Through these chapters, we aim to show how the incorporation of metadata and consideration of generative models pose concrete and interesting problems for interdisciplinary applied mathematics.

Outline of the Chapters

Each of the first two chapters offers a distinct perspective on node attributes. In Chapter 2, we study the adaptive voter model (AVM), a stochastic process in which node attributes are dynamically coupled to edge topology. Substantively, AVMs model the emergence of endogenous polarization in opinion-driven social networks. Despite their simplicity, AVMs have a surprisingly rich phenomenology, intensively studied by applied mathematicians and physicists over the last decade. This phenomenology

includes a family of quasistable manifolds governing the relationship between opinion densities and network topology. However, the coupling of node and edge states generates nonlinear low-dimensional approximations that often yield poor quantitative model predictions. We provide a new approximation scheme that exploits a physical asymmetry in certain parameter regimes, allowing the achievement of superior predictive performance from low-dimensional methods. These results allow us to accurately quantify the relationship between the decision-making of individual agents on the one hand and macroscopic levels of system polarization on the other. While this chapter does not explicitly focus on data science *per se*, it highlights the rich behavior that can be observed and studied when pure connectivity structure is combined with node attribute information.

In Chapter 3, we turn to the topic of spatial segregation. Segregation, like polarization, refers to systematic separation between agents of different attributes; unlike polarization, which is often modeled as a partially endogenous process, many forms of segregation reflect histories of state-sanctioned violence and oppression. Because of this, endogenous dynamical models are less relevant. On the other hand, we often have ample, high-resolution demographic data which can support sophisticated analytical techniques. However, as we argue in the chapter, extant techniques from the quantitative sociological literature often lack (a) mathematical foundations, (b) the ability to represent spatial proximity in principled ways, and (c) explicit representations of segregation at multiple spatial scales. Deploying the formalism of Bregman information geometry, we derive novel techniques for spatial analysis and visualization. These techniques employ maps of adjacency graphs to construct approximate manifold representations of spatially distributed demographic data. In contrast to the usual framing of node attributes as "metadata," in this case the node attributes constitute the primary data and therefore take center stage. We offer both macroscopic algorithms based on graph clustering methods and microscopic measures based on local information densities, proving a novel information-geometric identity in order to compute the latter quantity. Throughout this chapter, we emphasize the use of measures and concepts familiar within the quantitative segregation literature,

especially information decomposability. This emphasis guides us to construct domain-customized algorithms, especially the greedy agglomerative information maximization used to construct visualizations and information scaling curves.

Each of the final two chapters treats in some detail a problem in random graph null modeling. Many tasks in network data analysis depend on the comparison of one or more network statistics to their expectations or distributions under an appropriately-specified random graph null model. In general terms, a suitable null should incorporate "background features" of the data, but not the features directly under study. The selection of such a model is at least as much an art as a science, and it is usually necessary to reason about features of the data generating process in order to determine a reasonable null. That said, there are standard approaches. A common choice is "the configuration model," whose name we have placed in quotation marks because there are in fact several models which pass under this name in the greater network science community [76]. A shared feature of such models is the preservation of the first moments of the observed graph – the degree sequence. Typically, in the mathematical literature this preservation is required to be exact, while in the physics literature it is often required only in expectation.

In Chapter 5, we study the problem of estimating the expected adjacency matrix of the uniform distribution over multigraphs with a specified degree sequence. This distribution is a natural operationalization of the phrase "random network with fixed degree sequence," ubiquitous in the network science literature. However, common formulas that claim to calculate the expected adjacency matrix are based on an asymptotic connection to a different model (the stub-matching configuration model). The stub-matching configuration model, however, is misspecified on a wide class of common social data sets [76] and can lead to badly erroneous approximations for moments of the uniform model. Worse still, estimation via Monte Carlo sampling is not practical due to the large number of samples required and the conjectured poor mixing times of the associated Markov chains. We therefore seek analytical approximations. By treating a Monte Carlo sampler as a stochastic dynamical system, we are able to derive approximate equilibrium conditions, which we can solve to obtain

expressions with error bounds for key model moments in terms of an unknown vector $\boldsymbol{\beta}$. We provide a nonlinear equation that can be solved to approximate $\boldsymbol{\beta}$, and use a simple mountain pass argument to provide a qualified uniqueness theorem. Experiments show that the resulting approximation scheme reduces the error of standard methods by one and a half orders of magnitude, and significantly impacts the results of downstream data analysis. These results emphasize the need for analysts to carefully reason about whether the uniform or stub-matching configuration model is most appropriate for null-model comparison given the processes that generate their data.

Configuration models are traditionally defined on dyadic graphs, in which each edge links two nodes. Many modern data sets are *polyadic*, logging interactions between groups of arbitrary size. Example interactions including collaboration, communication, and copurchasing. While it is possible to project such data into dyadic graphs and model them using dyadic techniques, it is more faithful to the data generating processes to use natively polyadic techniques. Hypergraphs are natural representations of such data sets. In Chapter 4, we therefore generalize configuration models from random graphs to random hypergraphs. Our two configuration models preserve the two natural sets of first moments – the node-degree and edge-dimension sequences. We provide a unified Monte Carlo scheme for sampling from each. We then turn our attention to three data-analytic vignettes. In the first, we review classical results on clustering in social networks. In contrast to conventional wisdom, we find that several polyadic social networks are in fact *less* clustered than would be expected under a polyadic null model. We next study degree-assortativity, showing how polyadic null modeling can generalize and enrich classical measures. Finally, we illustrate how to leverage polyadic structure to define and test novel measures of higher-order edge correlations, including an asymptotic result that allows us to bypass explicit null model sampling for large polyadic data sets. Taken as a whole, these results highlight the importance of defining analytical methods that respect the processes by which network data is generated and collected, and extend the application of methods from random graph theory to network data.

Chapter 2

From Local Symmetries to Global Structure in Adaptive Voter Models

Adaptive voter models (AVMs) are simple mechanistic systems that model the emergence of mesoscopic structure from local networked processes driven by conflict and homophily. AVMs display rich behavior, including a phase transition from a fully-fragmented regime of "echo-chambers" to a regime of persistent disagreement governed by low-dimensional quasistable manifolds. Many extant methods for approximating the behavior of AVMs are either restricted in scope, expensive in computation, or inaccurate in predicting important statistics. In this work, we develop a novel, second-order moment closure approximation method for binary-state rewire-to-random and rewire-to-same model variants. We incorporate a small amount of noise via a random mutation term, which renders the system ergodic. Using ergodicity, we then approximate the voting process, which is non-Markovian in the second moments of the system, with a Markovian term near the phase transition. This approximation exploits an asymmetry between different classes of voting events. The resulting scheme enables us to predict the location of the phase transition and the active edge density in the regime of persistent disagreement, across the entire space of parameters and opinion densities. Numerically, our results are nearly exact for the rewire-to-random model, and competitive with extant approaches for the rewire-to-same model. Moreover, our computations display constant scaling in the mean degree, enabling approximations

for denser systems than previously possible. We conclude with suggestions for model refinements and extensions.

This work is based on the manuscript [50], forthcoming in the *SIAM Journal on Applied Mathematics*. This paper was written with mentorship, support, and collaboration by Peter J. Mucha, who served as the senior author. We are grateful to Feng (Bill) Shi for contributing code used for simulations, Hsuan-Wei Lee for contributing code used to construct the approximate master equation solutions shown in Figure 2-1, and Patrick Jaillet for helpful discussions. The mathematics, accompanying computations, and bulk of the discussions are my own.

2.1 Introduction

A common feature of social networks is trait-assortativity, the tendency of similar individuals to interact more intensely or frequently than dissimilar ones. Assortativity can be beneficial, allowing communities of individuals who share common beliefs or experiences to pursue shared goals. On the other hand, assortativity can also restrict flows of information and resources across heterogeneous populations. Recent scrutiny, for example, has fallen on the role of online platforms in promoting political polarization by allowing users to micromanage their contacts and information sources [10, 15].

The importance of trait assortativity has inspired various models of self-sorting populations. Among the most influential of these is the classical Schelling model [184], which models the emergence of spatial segregation through a preference of agents to live near a minimum number of similar neighbors. Inspired by this model, the authors of [99] consider the case of a social network in which agents are assigned an immutable attribute vector that may model demographics or opinions. Agents are allowed to destroy their connections to dissimilar partners and create new connections to similar ones, with the aversion to dissimilarity governed by a tunable parameter. The authors show that the model always generates segregated communities for any nonzero degree of dissimilarity aversion. Because the fixed node attributes are generated exogenously

to the system dynamics, this model is most appropriate for studying assortativity based on immutable or slowly-changing attributes, such as demographic variables. Contrasting to these dynamics is the family of voter models [56, 101], which are also defined on networks. In a typical voter model, each node is endowed with an opinion that evolves over time, usually via adoption of the opinion of a uniformly random neighbor. In original formulations, the network topology of a voter model is held fixed as opinions evolve.

In many networks, we naturally expect the opinions of individuals to both influence and be influenced by the connections they form. Over the past dozen years, a class of *adaptive network* models [204, 95, 97] has emerged to model such interacting influences. The distinguishing feature of these models is the dynamical coupling between node attributes and network topology. Such models have been studied in contexts including epidemic spreading [96, 132, 64, 124, 106] and strategic behavior [123, 159], but are most commonly deployed as models of opinion dynamics [103, 67, 223, 189, 131, 186, 165]. In this setting, they often appear as *adaptive* (or *coevolutionary*) *voter models* (AVMs). AVMs add opinion-based edge-rewiring to the opinion-adoption dynamics of the base voter model.[1] The tunable coupling of these processes generates polarized networks of opinion-based communities. AVMs thus constitute a class of "model organisms" [189] of endogenous fragmentation, polarization, and segregation in social and information networks.

Mathematically, AVMs display rich behavior, including metastability and phase transitions. However, the nonlinearity driving this rich behavior renders AVMs difficult to analyze even approximately. Many extant methods are restricted in scope, tractability, or accuracy, and often fail to provide insight into observed behaviors. Our aim in this article is to develop a class of approximation methods that both explain qualitative behaviors in these systems and provide analytical scope, computational efficiency, and predictive accuracy.

[1] Non-voter type updates are also possible in adaptive opinion-dynamics models; see e.g. [30] for a game-theoretic approach.

2.1.1 Outline of the Chapter

In Section 2.2, we formulate the class of binary-state AVMs studied here, review their behavior, and survey previous approaches developed for approximating their macroscopic behaviors. We study a model variant that includes a small amount of random opinion-switching ("mutation"), which renders the model ergodic. Using ergodicity, we develop in Section 2.3 an approximation scheme for the equilibrium macroscopic properties across the entirety of the model's phase space. Our scheme offers predictions for the point of emergence of persistent disagreement, which corresponds to the "fragmentation transition" in non-ergodic model variants. It also offers predictions for the density of disagreement once it emerges, including the arch-shaped quasistable manifolds characteristic of this class of models. We close in Section 5.6 with comparisons to the body of existing models, showing that we achieve favorable scope, accuracy, and computational complexity. Finally, we discuss promising extensions, both to our approximation methodology and to the model itself.

2.2 Adaptive Voter Models

An adaptive voter model is a first-order, discrete-time Markov process whose states are graphs with opinion-labeled nodes. Each state has the form $\mathcal{G} = (\mathcal{N}, \mathcal{L}, \mathcal{E})$, where $\mathcal{N}$ is a set of nodes and $\mathcal{E}$ a set of edges. We have $(u, v) \in \mathcal{E}$ means that an edge linking nodes u and v is present in $\mathcal{G}$. We denote by $\mathcal{N}(u)$ the neighborhood of u, comprising all nodes adjacent to u and u itself. The vector $\mathcal{L}$ maps $\mathcal{N} \to \mathcal{X}$ where $\mathcal{X}$ is an alphabet of possible states or opinions. We treat the node set $\mathcal{N}$ as fixed, while both $\mathcal{L}$ and $\mathcal{E}$ evolve stochastically in each time-step. We here restrict ourselves to the commonly-considered binary-state case, which we denote $\mathcal{X} = \{0, 1\}$, though multi-state variants [103, 186] are also of interest.

The temporal evolution of an AVM is characterized by superimposed voting dynamics on $\mathcal{L}$ and edge-rewiring dynamics on $\mathcal{E}$. To these, our model variant adds a third process in the form of random opinion switching or "mutation" in $\mathcal{L}$. We specify the discrete-time stochastic dynamics $(\mathcal{E}(t), \mathcal{L}(t)) \mapsto (\mathcal{E}(t+1), \mathcal{L}(t+1))$ as follows:

1. With probability $\lambda \in [0,1]$, **mutate**: uniformly sample a node $u \in \mathcal{N}$ and set $\mathcal{L}_u(t+1) \leftarrow \text{uniformChoice}(\mathcal{X} \setminus \{\mathcal{L}_u(t)\})$. Note that mutation does not add states to the opinion alphabet $\mathcal{X}$, which is fixed. In the binary-state case, a mutation step deterministically maps $\mathcal{L}_u(t+1) \leftarrow 1 - \mathcal{L}_u(t)$.

2. Otherwise (with probability $1-\lambda$), sample an edge $(u,v) \in \mathcal{E}(t)$ uniformly from the set $\{(u,v) : \mathcal{L}_u(t) \neq \mathcal{L}_v(t)\}$ of *active* edges (also referred to in some studies as *discordant* edges). The orientation of (u,v) is uniformly random. Then,

   (a) With probability $\alpha \in [0,1]$, **rewire**: delete the (undirected) edge (u,v) and add edge (u,w) selected according to one of the following two rules depending on the model variant being used. In the *rewire-to-random* model variant, w is chosen uniformly from $\mathcal{N} \setminus \mathcal{N}(u)$. In the *rewire-to-same* variant, w is chosen uniformly from the set $S_u = \{w \in \mathcal{N} \setminus \mathcal{N}(u) | \mathcal{L}_w(t) = \mathcal{L}_u(t)\}$. In the rewire-to-same case, it may in principle occur that u is already connected to all members of the set S_u. In this case, we simply pass to the next iteration, starting from Step 1, without modifying $\mathcal{G}$.

   (b) Otherwise (with probability $1-\alpha$) **vote**: $\mathcal{L}_u(t+1) \leftarrow \mathcal{L}_v(t)$.

From a modeling perspective, mutation may represent phenomena such as media influence, noisy communication, or finite agential memory. The mutation mechanism is reminiscent of the "noisy" voter model of [91], and was introduced in an adaptive model variant by [110].

The rewiring and voting steps both occur after sampling an active edge uniformly at random. Other sampling schemes are also possible. The sampling in [103], for example, selects a uniformly random node u with nonzero degree. Then, a uniformly random neighbor v of u is chosen. Rewiring occurs with probability α and voting with probability $1-\alpha$ regardless of their respective opinions. In the model introduced by [204] and further studied by [201, 110, 116], u and v are chosen similarly, but in the event that $\mathcal{L}_u(t) = \mathcal{L}_v(t)$ nothing happens and the sampling step is repeated. Sampling via active edges as we do here was to our knowledge introduced in [67] and employed in many recent studies [63, 35, 36, 189, 18, 177]. The authors of [67]

note that models with different sampling mechanisms nevertheless display similar qualitative – and often quantitative – macroscopic behaviors.

AVMs are usually studied through a standard set of summary statistics. Let $n = |\mathcal{N}|$ be the number of nodes, $m = |\mathcal{E}(t)|$ the number of edges, and $c = 2m/n$ the mean degree. Since the dynamics conserve n and m, c is time-independent and may be regarded as an additional system parameter. Let $N_i(t) = |\{u \in \mathcal{N} \mid \mathcal{L}_u(t) = i\}|$ be the number of nodes holding opinion i at time t. Let $\mathbf{q}(t) = (q_0(t), q_1(t)) = n^{-1}(N_0(t), N_1(t))$ be the vector of opinion densities. For each pair i and j of opinions in $\mathcal{X}$, let $M_{ij}(t) = |\{(u,v) \in \mathcal{E} \mid \mathcal{L}_u(t) = i,\ \mathcal{L}_v(t) = j\}|$ be the number of *oriented* edges between nodes of opinion i and nodes of opinion j. Note that $M_{ij}(t) = M_{ji}(t)$ and $\sum_{i,j\in\mathcal{X}} M_{ij}(t) = 2m$ at all times t, since each (undirected) edge is counted twice in the vector $\mathbf{M}$, once in each of two orientations. Let $\mathbf{X}(t) = (X_{00}, X_{01}, X_{10}, X_{11}) = (2m)^{-1}\mathbf{M} = (2m)^{-1}(M_{00}(t), M_{01}(t), M_{10}(t), M_{11}(t))$ be the vector of *oriented* edge densities. We define the scalar $R(t) = X_{01}(t) + X_{10}(t) = 2X_{01}(t)$ to be the overall density of active edges. By construction, $R(t)$ is a random variable on the interval $[0,1]$. Let $\mathbf{x}(t) = \mathbb{E}[\mathbf{X}(t)]$ and $\rho(t) = \mathbb{E}[R(t)]$, with expectations taken with respect to the time-dependent measure of the Markov process. Note that the objects $\mathcal{L}(t)$, $\mathbf{X}(t)$, and $R(t)$ are random functions of time t, while $\mathbf{x}(t)$ and $\rho(t)$ are deterministic functions of time.

Most previous studies have considered AVM variants without mutation, corresponding in our setting to $\lambda = 0$. In this setting, any state in which $R = 0$ is an absorbing state of the Markov chain. Such a state consists of one or more connected components within each of which consensus reigns. Letting $C(u)$ denote the connected component of node u in the absorbing state in this regime, it holds that $C(u) = C(v)$ implies $\mathcal{L}_u = \mathcal{L}_v$ for any nodes u and v. As discussed in both [103] and [67], there is a phase transition in the (random) final value $\mathbf{q}^*$ of the opinion densities in the absorbing state. In both model variants, there is a critical value α^*, depending on $\mathbf{q}(0)$, such that, if $\alpha \geq \alpha^*(\mathbf{q}(0))$, $|\mathbf{q}^* - \mathbf{q}(0)|_1 = O\left(\frac{\log n}{n}\right)$ with high probability as n grows large. Thus, in the large n limit, the opinion densities are not appreciably altered by the dynamics. We refer to this as the "subcritical" parameter regime. On the

other hand, if $\alpha < \alpha^*(\mathbf{q}(0))$, $\mathbf{q}^*$ is governed by a bimodal distribution parameterized by α, c, and the model rewiring variant. In both models, the phase transition marks the point at which the voting dynamics outstrip the rewiring dynamics, in the sense that the rewiring dynamics are no longer fast enough to resolve most disagreements, and therefore also corresponds to a transition in the time to reach the final state [103, 177]. We refer to this regime as "supercritical." Note that in the $\lambda = 0$ case, the non-ergodicity of the model implies that all these results are to an extent dependent on the initial state $\mathcal{G}_0$ of the AVM. While this dependence is generally weak in the standard AVM we consider here, in related model variants [120] the initial conditions may often dominate even the rewiring mechanism in determining the final system state.

In [67], the authors show via simulation and analytical methods that the same phase transition marks the emergence of a *quasistable manifold* along which the system dynamics evolve. This manifold is well-approximated by a concave parabola in the (q_1, ρ)-plane, reflected by its colloquial name, "the arch." Similar arches were observed for an AVM variant in [204] and for a non-adaptive voter model in [203]. When $\alpha > \alpha^*(\mathbf{q}(0))$, ρ converges rapidly to 0 while $\mathbf{q}$ remains nearly constant. When $\alpha < \alpha^*$, on the other hand, the trajectory converges to a point on the arch, and then slowly diffuses along it until reaching an absorbing state at one of the two bases. In the rewire-to-random arch, α^* depends on q_1, and the arch is therefore supported on a proper sub-interval of $[0, 1]$. In contrast, the rewire-to-same transition is independent of q_1, and the associated arch is supported on the entirety of $[0, 1]$. The bases of the arch correspond to the modes in the long-run distribution of $\mathbf{q}^*$.

While multiple studies have achieved insight via numerical study of simulation traces [215, 186, 110], analytical insight into the phenomenology of AVMs remains limited. The central analytical project is to estimate the behavior of the expected edge-density ρ as a function of the parameters λ, α, and c, as well as the opinion density $\mathbf{q}$.[2] The most modest task is to estimate the phase transition α^* in the

[2]Recent papers have studied other features of interest, such as approximate conservation laws [201] and network topology near the phase transition α^* [106]; however, we will not pursue these themes further.

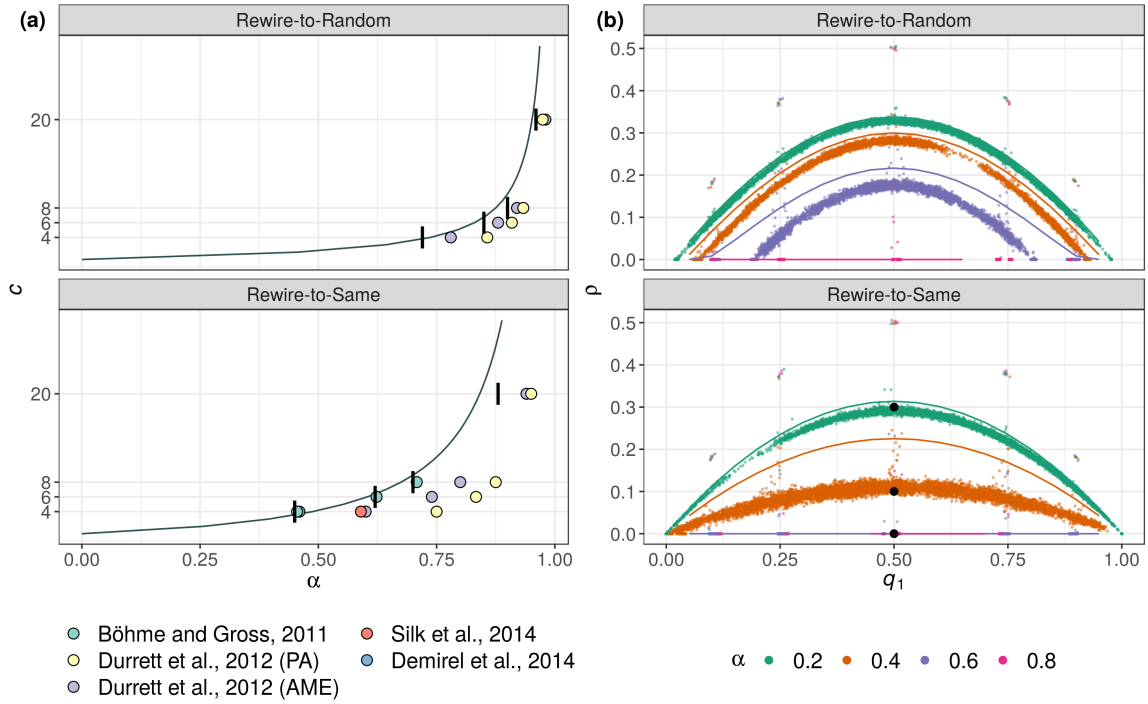


Figure 2-1: Overview of phase transitions, quasistable manifolds, and previous estimation methods for AVMs.
(a): Estimates of the phase transition α^* in the expected density ρ of active edges when $\mathbf{q}(0) = \left(\frac{1}{2}, \frac{1}{2}\right)$, for varying mean degree c. Vertical black lines give the empirical location of the phase transition, determined numerically as the smallest value of α for which $\rho > 0.01$. Colored points give estimates of the phase transition from extant methods. The $c = 4$ estimates for the rewire-to-same variant of Böhme and Gross [35] and Demirel et al. [63] overlap. The solid line gives the estimate of our proposed method, obtained by solving Equation (2.11). (b): Quasi-stable arches in the (q_1, ρ)-plane for varying α. Points are sampled from simulations at intervals of $5,000$ time-steps. Black dots on the rewire-to-same panel give the active-motif estimate of [63] for the symmetric top of the arch. The solid lines give the approximate master equation estimates of [67]. Simulations in this and subsequent figures were performed with $N = 10^4$ nodes and mutation rate $\lambda = 2^{-10} \approx 10^{-3}$. All simulations were initialized with an Erdös-Rényi $G(n, p)$ graph on $n = 10^4$ nodes with specified mean degree c. Each node independently chooses its initial opinion 0 or 1 with equal probability. We performed 10^7 total simulation steps, and sampled the process after a burn-in period of 2×10^6 steps. This process was repeated ten times for each combination of parameters α, and c.

case of symmetric opinion initialization $\mathbf{q}(0) = \left(\frac{1}{2}, \frac{1}{2}\right)$. Figure 2-1(a) summarizes a selection of extant methods to approximate the location of the phase transition in these model variants over the last decade and compares them to the observed emergence of the top of the arch in model simulations. The pair approximation (PA) [116, 67] is an all-purpose method for binary-state models that usually produces qualitatively correct but quantitatively poor results. Indeed, Figure 2-1 shows that the pair approximation overestimates the location of the phase transition, performing especially poorly in the rewire-to-same model variant. More specialized methods are required to obtain quantitatively reasonable estimates. The method of [35] uses compartmental equations to accurately estimate the rewire-to-same phase transition with symmetric opinion densities, finding close agreement with observation in this restricted task. In [18] the authors apply stopping-time arguments to give a rigorous proof of the existence of a phase transition in both model variants. However, their results apply only in the context of dense limiting graphs and do not explicitly predict the value of α^*.

Other schemes provide estimates not only of the transition but also of the quasistable supercritical active link density ρ when $\mathbf{q}(0) = \left(\frac{1}{2}, \frac{1}{2}\right)$. The authors of [63] propose a compartmental approach based on *active motifs* to estimate the phase transition and arch in the symmetric opinion rewire-to-same model variant. An active motif consists of a node and a number of active links attached to it; a system of ordinary differential equations may be obtained by approximately tracking the evolution of active motif densities in continuous time. The resulting estimate of the phase transition (Figure 2-1(a)) and of the top of the arch (Figure 2-1(b)) are both highly accurate, but require an active-link localization assumption specific to the rewire-to-same variant. The authors of [189] follow a related approach for the rewire-to-same variant based on more general *active neighborhoods*. Active neighborhoods count the numbers of both active and inactive links attached to a given node. They obtain an analytic approximation by transforming the resulting system into a single partial differential equation governing the generating functions of the neighborhood densities. The resulting estimate of the phase transition (fig. 2-1(a)) and the active link

density (not shown) are, however, uniformly dominated in accuracy by the explicit active-motif approach.

To our knowledge, the only methods for approximating the complete arch are the pair approximation and the approximate master equations (AMEs, [88]) used in [67]. Approximate master equations are similar to active-neighborhood techniques, but are formulated explicitly for the case of general opinion densities $\mathbf{q}$. For small mean degree, approximate master equations can provide relatively accurate predictions of the rewire-to-random phase transition (Figure 2-1(a)) and qualitatively reasonable estimates of the arches (Figure 2-1(b)), though the shapes of the arches may be somewhat distorted. Their estimates for α^* and ρ in the rewire-to-same variant are substantially worse, although the qualitative shape of the arches appears correct. AMEs are constrained by their computational cost: to obtain a solution requires the numerical solution of $\Theta(k_{\max}^2)$ coupled differential equations, where $k_{\max}$ is the largest node-degree expected to emerge in the course of a simulation, and therefore depends at least linearly on the mean degree c. The scheme thus rapidly becomes impractical for high enough mean degree or for initially skewed degree distributions.

2.2.1 AVMs with Mutation

The approximation scheme we will develop in Section 2.3 depends on the presence of mutation in the model, i.e. $\lambda > 0$. The introduction of mutation has an important technical consequence: the process is ergodic, up to symmetry.

Definition 1. *A* labeled graph isomorphism *of a state $\mathcal{G} = (\mathcal{N}, \mathcal{L}, \mathcal{E})$ is a permutation $\tau : \mathcal{N} \to \mathcal{N}$ such that $(u, v) \in \mathcal{E}$ iff $(\tau(u), \tau(v)) \in \mathcal{E}$ and $\mathcal{L}_u = \mathcal{L}_{\tau(u)}$ for all $u \in \mathcal{N}$. We write $\overline{\mathcal{G}}$ for the equivalence class of $\mathcal{G}$ under labeled graph isomorphisms.*

Theorem 1. *When $\lambda > 0$, if $\binom{n-4}{2} \geq m - 1$, the process $\overline{\mathcal{G}}(t)$ is ergodic.*

Proof. We will first show aperiodicity by constructing cycles of lengths 2 and 3 in the state space. To construct a cycle of length 2, simply choose a node and perform two sequential mutation steps. The construction of a cycle of length 3 is slightly more involved. Pick an edge $e \in \mathcal{E}$. Label one end u, and the other end v_1. Pick two more

nodes v_2 and v_3. Using mutation and rewiring steps, remove all edges connected to u, v_1, v_2, and v_3 except for e. This can always be done by hypothesis, since the remaining $m-1$ edges may be placed among the $\binom{n-4}{2}$ pairs of remaining nodes via mutation and rewiring steps. Using mutation steps, set $\mathcal{L}_u = \mathcal{L}_{v_2} = 0$ and $\mathcal{L}_{v_1} = \mathcal{L}_{v_3} = 1$. Call this initial state $\mathcal{G}$. Then, consider the following sequence:

1. Rewire $(u, v_1) \mapsto (u, v_2)$.
2. Mutate $\mathcal{L}_{v_2} \leftarrow 1$.
3. Mutate $\mathcal{L}_{v_1} \leftarrow 0$.

Call the end state $\mathcal{G}'$. Each of these steps is supported in both rewire-to-same and rewire-to-random model variants. Furthermore, the permutation τ that interchanges v_1 and v_2 is a labeled isomorphism from $\mathcal{G}$ to $\mathcal{G}'$. We have therefore constructed a supported cycle of length 3 in the state space of the process $\overline{\mathcal{G}}(t)$, completing the proof of aperiodicity.

To show irreducibility, let $\mathcal{G}_1 = (\mathcal{N}, \mathcal{L}_1, \mathcal{E}_1)$ and $\mathcal{G}_2 = (\mathcal{N}, \mathcal{L}_2, \mathcal{E}_2)$ be elements of the state space of a single AVM. Since $|\mathcal{E}_1| = |\mathcal{E}_2| = m$, we have $|\mathcal{E}_1 \setminus \mathcal{E}_2| = |\mathcal{E}_2 \setminus \mathcal{E}_1|$. These sets may therefore be placed in bijective correspondence. For each edge $e = (u, v) \in \mathcal{E}_1 \setminus \mathcal{E}_2$, we arbitrarily identify $e' = (u', v') \in \mathcal{E}_2 \setminus \mathcal{E}_1$. Perform the sequence of rewirings $(u, v) \mapsto (u, v') \mapsto (u', v')$ possibly with mutation steps in order to activate the edges as needed. Doing so reduces the set $\mathcal{E}_1 \setminus \mathcal{E}_2$ by one edge. Repeat this process inductively until $\mathcal{E}_1 \setminus \mathcal{E}_2 = \varnothing$; that is, until $\mathcal{E}_1 = \mathcal{E}_2$. Finally, perform mutation steps on all nodes u on which $\mathcal{L}_1$ and $\mathcal{L}_2$ disagree. The result is a path of nonzero probability through the state space of $\mathcal{G}(t)$ and therefore of $\overline{\mathcal{G}}(t)$, as was to be shown. □

Since the process $\overline{\mathcal{G}}(t)$ is ergodic, it possesses an equilibrium measure η supported on the entirety of its state space. In the remainder of this paper, we will abuse notation by identifying $\mathcal{G}$ with $\bar{\mathcal{G}}$ and referring to η as the equilibrium distribution of $\mathcal{G}(t)$. An important consequence of ergodicity is that all properties of η derived for the mutating AVM are independent of the state in which the network is initialized. This situation contrasts with the extant literature discussed previously (with the exception

of [110]), in which the model behavior can, in principle, depend on initialization. Ergodicity also implies that states with $R = 0$ are no longer absorbing. Instead, a typical sample from η displays bifurcated structure closely aligned with the opinion groups, with dense connections between common opinions and sparser connections between differing opinions. This behavior of the mutating AVM thus makes it a more flexible model of social processes in which long-standing disagreement may influence connections. We focus on the limit of small λ, which allows us to derive approximations for the non-mutating AVMs. In particular, the equilibrium measure η concentrates around the $\lambda = 0$ arch, allowing us to describe the arch as the expected active link density $\rho^* = \mathbb{E}_\eta[R]$ under the equilibrium measure η.

2.3 Model Analysis

Our strategy is to study perturbations from the fully-fragmented state $R = 0$. These perturbations are induced by mutation, without which the fully-fragmented state is absorbing. While many existing techniques amount to continuous-time mass-action laws for system moments, our methods are fundamentally discrete and local in that we study changes in the edge density vector $\mathbf{X}$ stemming from a single mutation event. Carefully-chosen approximations in this regime can be expected to be accurate near the critical point.

Assume that λ is small but positive. Suppose that at time t, $R = 0$. In this state, $\mathcal{G}(t)$ consists of one or more connected components within which the opinion function $\mathcal{L}$ is constant. Suppose now that, at time $t + 1$, node u on component $C(u)$ changes its opinion from 0 to 1 through mutation. Because opinions on $C(u)$ are otherwise uniform, all active links present in component $C(u)$ are contained in the neighborhood of u itself. In particular, any additional active links that may be generated over short timescales near u, as measured by the geodesic distance on $\mathcal{G}$.

Let T be the hitting time of the event $R = 0$; i.e., the amount of time required to return to the fully-fragmented state. We can distinguish two regimes, depending on the scaling of $\mathbb{E}[T]$ with n.

1. **Subcritical**: We have $\mathbb{E}[T] = O(1)$. Intuitively, this occurs when u's dissenting opinion is either snuffed out by voting events or "quarantined" by rewiring events in a small number of time steps. This case always occurs when $\alpha = 1$, since T is then simply the time until each active link has been rendered inactive via rewiring. The expected number of active edges scales as $n\rho^* = O(1)$, since there are only $O(1)$ time steps in which additional active edges may be generated. We therefore have $\rho^* \to 0$ as n grows large.

2. **Supercritical**: We have $\mathbb{E}[T] = O(n^2)$, corresponding to the consensus-time of the non-adaptive voter model [101]; as such, this case always occurs for $\alpha = 0$. Mechanistically, u's dissenting opinion triggers a cascade of active edge-generation through voting and rewiring events with nonzero probability. In this case, the number R of active edges scales with n (see, e.g. [203]), and the equilibrium active edge density ρ^* is nonzero as n grows large.

These two regimes are separated by critical values in the parameters α, λ, and c. Indeed, the transition in α is precisely that described previously for the $\lambda = 0$ case. The situation is thus reminiscent of the standard Galton-Watson branching process [14], in which the criticality of the aggregate process can be characterized locally by the reproductive potential of a single node.

To develop quantitative approximations, we therefore study the local dynamics around node u. At the moment that node u changes its opinion from 0 to 1, all nodes on $C(u)$ other than u itself have opinion 0. Even if further mutation events take place, it will still be true that local neighborhoods of nodes in $C(u)$ are statistically dominated by opinion 0 nodes. Similarly, we can distinguish a local minority — initially comprising node u alone — of opinion 1. In the subcritical regime, every connected component possesses a clearly defined local minority and local majority. In the supercritical regime, these distinctions degrade as the number of active links increases.

We will use this physical intuition to formulate a closed-form approximation in the neighborhood of the critical point. Then, the dynamics in the expected edge counts

may be written

$$\mathbf{m}(t+1) - \mathbf{m}(t) = \mathbb{E}\left[\lambda \mathbf{W}(\mathcal{G}(t)) + (1-\lambda)\alpha \mathbf{R}(\mathcal{G}(t)) + (1-\lambda)(1-\alpha)\mathbf{V}(\mathcal{G}(t))\right] , \quad (2.1)$$

where $\mathbf{W}$, $\mathbf{R}$, and $\mathbf{V}$ are (random) functions of the graph state $\mathcal{G}(t)$ giving the increments in $\mathbf{m}$ due to mutation, rewiring, and voting, respectively. Importantly, $\mathbf{W}$ and $\mathbf{R}$ depend only on $\mathbf{q}$ and $\mathbf{x}$, the expected first and second moments of $\mathcal{L}$. Starting with the former, the entries of the expected mutation term may be written

$$\mathbb{E}[\mathbf{W}(\mathcal{G})] = \mathbf{w}(\mathbf{x}) = c \begin{bmatrix} x_{10} - x_{00} \\ x_{00} - x_{10} + x_{11} - x_{01} \\ x_{00} - x_{10} + x_{11} - x_{01} \\ x_{01} - x_{11} \end{bmatrix} . \quad (2.2)$$

We illustrate by deriving the expression for $\mathbf{w}_{00}(\mathbf{x})$. Edges between nodes of opinion 0 are created when an opinion-1 node on an active edge mutates. A uniformly random opinion-1 node has in expectation cx_{10} active edges available to transform into 0-0 edges upon mutation. Similarly, 0-0 edges are destroyed when one of the incident nodes mutates. A uniformly random opinion-0 node has in expectation cx_{00} inactive edges that are destroyed upon mutation. The expressions for the other entries of $\mathbf{w}$ are derived by parallel arguments. The rewiring terms $\mathbf{r}$ are written as follows:

$$\mathbb{E}[\mathbf{R}(\mathcal{G})] = \mathbf{r}(\mathbf{q}) = \begin{cases} \left[q_0, -\frac{1}{2}, -\frac{1}{2}, q_1\right]^T & \text{rewire-to-random} \\ [1, -1, -1, 1]^T & \text{rewire-to-same.} \end{cases}$$

Notably, the rewiring function depends on the opinion densities $\mathbf{q}$ only in the rewire-to-random case, because the rewire-to-same variant always removes exactly one active edge, replacing it with an inactive one in a rewiring step. To derive the expression for the rewire-to-random case, we can condition on the opinion of the node that "keeps" the active edge e. If the 0-opinion node keeps e, then, with probability q_0, e joins to another opinion 0 node, destroying the active edge and creating a 0-0 edge. A similar

argument accounts for the q_1 term. Summing up the ways for an active edge to be removed, we have $r_{01}(\mathbf{q}) = -\frac{1}{2}(q_0 + q_1) = -\frac{1}{2}$, as was to be shown.

The computations above show that the mutation and rewiring dynamics in $\mathbf{X}$ are Markovian: for any fixed $\mathbf{q}$, when $\alpha = 1$, $\mathbf{X}$ is a Markov process. Because of this, computing $\mathbf{x}$ in the $\alpha = 1$ case for fixed $\mathbf{q}$ reduces to solving a four-dimensional linear system subject to nonnegativity and normalization constraints. Unfortunately, the voting term $\mathbf{v}(\mathcal{G}(t)) = \mathbb{E}[\mathbf{V}(\mathcal{G})]$ cannot be similarly parsed in terms of $\mathbf{q}$ and $\mathbf{X}$, because the voting dynamics depend on higher graph moments and are therefore non-Markovian in these variables. We may therefore view the short-timescale dynamics of $\mathbf{X}$ for fixed $\mathbf{q}$ as a mixture of Markovian opinion-switching and rewiring processes with a non-Markovian voting process. Our strategy is to approximate the expectation of the non-Markovian voting term with a Markovian approximation near the phase transition, using the asymmetry between local minorities and majorities. This approximation supposes that $\mathbb{E}[\mathbf{V}(\mathcal{G}(t))] \approx \hat{\mathbf{v}}(\mathbf{q}, \mathbf{x})$ near the critical regime, with the function $\hat{\mathbf{v}}$ of $\mathbf{q}$ and $\mathbf{x}$ to be determined. Note that doing so and setting the lefthand side of eq. (2.1) equal to zero removes all dependence on the time-step t except dependence through $\mathbf{q}$ and $\mathbf{x}$. We therefore suppress the argument t throughout the remainder of this section. All expectations are taken with respect to the stationary distribution η.

To construct $\hat{\mathbf{v}}$, we study the local neighborhood of a node u that has just changed its opinion from $\bar{\imath} \in \{0, 1\}$ to $i \in \{1, 0\}$. We denote expectations conditioned on this event using the shorthand $\mathbb{E}[\cdot|i]$. Immediately after this event, u possesses an initial random number J_0 of inactive and K_0 of active edges. The distributions of J_0 and K_0 depend on $\mathbf{q}$, $\mathbf{x}$, c, and their moments, as well as the conditions under which node u changed its opinion. If u changes its opinion due to a mutation on an otherwise constant-opinion component, then $J_0 = 0$. On the other hand, if u changes its opinion through a voting event, then $J_0 \geq 1$, since there must have been a node to pass on the opinion to u. To compute $\hat{\mathbf{v}}$, we track each of the K_0 active edges until each of them has been rendered inactive, counting voting events along the way.

Under timescale-separation and mean-field assumptions, these calculations may be

carried out in closed form. The assumption of timescale-separation supposes that $\mathcal{G}$ changes slowly relative to the neighborhood of node u, so that only update steps that sample the initial K_0 edges require accounting. The mean-field assumption supposes that nodes in the local majority have degree distributions governed by the global network average $\mathbf{x}$, reflecting the fact that, by definition, most nodes are members of their respective local majorities. These assumptions are approximately correct when the active edge density R and mutation rate λ are both small, and will tend to degrade when either quantity is increased.

Define the vector $\mathbf{c}$ with components $c_{ij} = cx_{ij}/q_i$. Each entry c_{ij} gives the average number of neighbors of type j of a node of type i. Note that, though $x_{ij} = x_{ji}$, it is not generally the case that $c_{ij} = c_{ji}$ unless $\mathbf{q} = \left(\frac{1}{2}, \frac{1}{2}\right)$. The random variable K_0 is the number of opinion $\bar{\imath}$ neighbors incident to u immediately prior to u changing opinion; under the mean-field assumption, we therefore have $\mathbb{E}[K_0|i] = c_{\bar{\imath}\bar{\imath}}$. Meanwhile, $\mathbb{E}[J_0|i] = 1 + c_{\bar{\imath}i}$ if u changed its opinion due to voting and $\mathbb{E}[J_0|i] = 0$ if u changed its opinion due to mutation. Since we assume λ to be small and mutations to therefore be slow, we focus on the former case.

We need to track multiple types of voting events, and we define random variables for each.

1. Neighbors of u may vote. By the assumption of timescale-separation, each such vote occurs along one of the K_0 initial active edges. Let E denote the (random) number of such votes.

2. Nodes not attached to u may vote. In the rewire-to-random model, such events may occur after an active edge attached to u is rewired away from u but remains active, allowing for a later time at which one of the two nodes on this edge votes to render the edge inactive. In the rewire-to-same model variant, this type of voting event does not occur. Let F denote the (random) number of such voting events.

3. Node u itself may vote prior to all of its K_0 active edges becoming inactive or removed from u. Let G denote the indicator random variable for this event.

We next write down vectors tracking the impact of each of the above voting event types on $\mathbf{m}$, the vector of expected global edge counts. We first compute the vector $\mathbf{e}_i(\mathbf{c})$ whose entries give the expected increment in $\mathbf{m}$ due to a Type 1 event. Since votes occur along active edges, a Type 1 event consists of a neighboring node v changing opinion from $\mathcal{L}_v = \bar{\imath}$ to $\mathcal{L}_v = i$. In this event, edge (u, v) is rendered inactive. At node v, $c_{\bar{\imath}\bar{\imath}}$ edges are activated in expectation, and $c_{\bar{\imath}i}$ edges are rendered inactive as i-i edges. Both of these expressions are implied by the mean-field approximation. We therefore have

$$\begin{aligned}\mathbf{e}_i(\mathbf{c}) &= \frac{1}{2}\left(-2\mathbb{E}[K_0|i], \mathbb{E}[K_0|i] - \mathbb{E}[J_0|i], \mathbb{E}[K_0|i] - \mathbb{E}[J_0|i], 2\mathbb{E}[J_0|i]\right)\right) \\ &= \frac{1}{2}\left(-2c_{\bar{\imath}\bar{\imath}}, c_{\bar{\imath}\bar{\imath}} - c_{\bar{\imath}i} - 1, c_{\bar{\imath}\bar{\imath}} - c_{\bar{\imath}i} - 1, 2(1 + c_{\bar{\imath}i})\right) \ . \end{aligned} \tag{2.3}$$

The expected increment vector for Type 2 events may again be computed via the mean-field approximation. Since the edges involved in Type 2 events are not connected to u, the increment is independent of $\mathcal{L}_u$. We therefore have

$$\mathbf{f}(\mathbf{c}) = \frac{\mathbf{e}_0(\mathbf{c}) + \mathbf{e}_1(\mathbf{c})}{2} \ . \tag{2.4}$$

The analysis for Type 3 events is more subtle. For $i = 1$, this term has components

$$\mathbf{g}_1(\mathbf{q}, \mathbf{c}) = \frac{1}{2}\left(2\mathbb{E}[GK|1], \mathbb{E}[G(J-K)|1], \mathbb{E}[G(J-K)|1], -2\mathbb{E}[GJ|1]\right) \ , \tag{2.5}$$

where J (respectively, K) are the number of inactive (active) edges attached to u at the time of voting, and G is the event that u votes prior to deactivation.

To complete the approximation scheme, it is necessary to first compute the expectations appearing in Equation (2.5) and then compute the expected number of events of each type. We begin with K, the active edge count at the time that u votes. Conditioned on a fixed initial number K_0 of active edges and u's opinion i, K

is distributed as a truncated geometric:

$$\mathbb{P}(K = k|i, K_0) = \begin{cases} (1-\beta_i)\beta_i^{K_0-k} & 1 \le k \le K_0 \\ \beta_i^{K_0} & k = 0, \end{cases}$$

where β_i is the probability that an event is not a vote by u, given that it removes a discordant edge from u and that u has opinion i. This probability is given explicitly by

$$\beta_i = \begin{cases} \frac{1+\alpha q_i}{2-\alpha(1-q_i)} & \text{rewire to random} \\ \frac{1+\alpha}{2} & \text{rewire to same.} \end{cases} \tag{2.6}$$

To derive the rewire-to-random expression, we enumerate the events that remove an active edge (u, v) from u, given that (u, v) is sampled for update. A vote by either node u or node v deactivates the edge, and occurs with probability $1-\alpha$. A rewiring event in which v maintains the edge removes the edge from u and occurs with probability $\alpha/2$. A rewiring event in which u maintains the edge occurs with probability $\alpha/2$, and deactivates the edge with probability q_i in the rewire-to-random case. The total rate of active edge removal from u is therefore $2-\alpha(1-q_i)$. The rate of active edge removal, excluding Type 3 voting events, is $2-\alpha(1-q_i)-(1-\alpha) = 1+\alpha q_i$. A similar derivation yields the expression for the rewire-to-same variant.

The probability of u voting prior to deactivation, conditioned on K_0, is

$$\mathbb{E}[G|K_0, i] = \mathbb{P}(K \ge 1) = 1 - \beta_i^{K_0} .$$

Averaging over K_0 yields

$$\mathbb{E}[G|i] = \sum_{k_0} \mathbb{P}(K_0 = k_0)(1 - \beta_i^{k_0}) = 1 - \phi_{K_0}(\beta_i) ,$$

where $\phi_{K_0}(z) = \sum_{k=1}^{\infty} \mathbb{P}(K_0 = k) z^k$ is the probability generating function of K_0. Some previous work (e.g. [204]) explicitly models quantities such as K_0 as binomial or

Poisson random variables. In our experiments, the crude approximation $\mathbb{E}[G|i] \approx 1 - \beta_i^{\mathbb{E}[K_0|i]} = 1 - \beta_i^{c_{\overline{ii}}}$ yields similar results with much faster computations, and is therefore used in the results presented below.

The expected number of active edges at the time that u votes is

$$\begin{aligned}\mathbb{E}[GK|i] &= \mathbb{E}_{K_0}\mathbb{E}[GK|i, K_0] \\ &= \mathbb{E}_{K_0}\left[K_0 - \frac{\beta_i(1-\beta_i^{K_0})}{1-\beta_i}\middle| i\right] \\ &= \mathbb{E}[K_0|i] - \frac{\beta_i}{1-\beta_i}\mathbb{E}[G|i]\,.\end{aligned}$$

This accounts for the decay in the local active edge density around u. It remains to compute $\mathbb{E}[E|i]$, $\mathbb{E}[F|i]$, and $\mathbb{E}[GJ|i]$. To do so, it is useful to introduce the coefficients

$$\varepsilon_i = \begin{cases} \frac{1-\alpha(1-q_i)}{1+\alpha q_i}\,, \\ \frac{1}{1+\alpha}\,, \end{cases} \qquad \sigma_i = \begin{cases} q_1\frac{2(1-\alpha)}{2-\alpha}\,, & \text{rewire to random}\,, \\ 0\,, & \text{rewire to same.} \end{cases} \tag{2.7}$$

The coefficient ε_i gives the probability that an event that removes an active edge from u, other than a vote by u, produces an inactive edge either through rewiring or through a vote by a neighbor of u. The coefficient σ_i gives the probability that an active edge which is rewired but not immediately deactivated is ultimately deactivated via a voting event. The derivations of these coefficients are similar to that of β_i above.

Node u begins with an initial number J_0 of inactive edges, and gains more via rewiring and voting. At the time that u votes, in expectation $\mathbb{E}[K_0|i] - \mathbb{E}[GK|i]$ active links have been removed; each has a probability ε_i of being deactivated while remaining attached to u. The expected number of inactive edges at the time that u votes is therefore

$$\mathbb{E}[GJ|i] = \mathbb{E}[J_0|i] + \varepsilon_i\left(\mathbb{E}[K_0|i] - \mathbb{E}[GK|i]\right)\,.$$

To compute $\mathbb{E}[E|i]$, the expected number of Type 1 events, we note that a voting

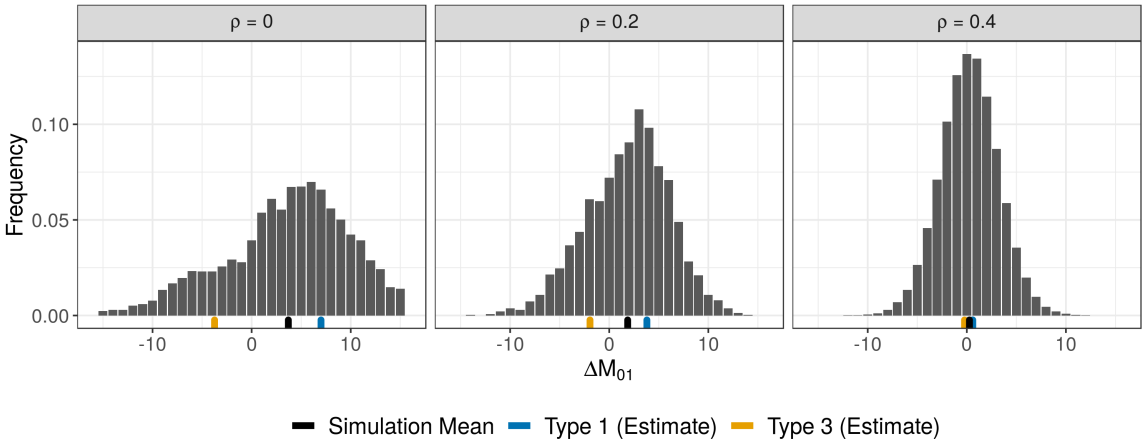


Figure 2-2: Illustration of the asymmetry between Type 1 and Type 3 events. Histograms give the impact of a voting event on M_{01}, the number of active edges. Each panel corresponds to a different value of the expected active edge density ρ. The expected impact of Type 1 (blue) and Type 3 (orange) events are shown in the horizontal margin, as well as the simulation mean (black). Simulations performed on a rewire-to-random AVM of $n = 10^4$ nodes and $c = 8$, with varying rewiring rate α under the conditions described in Figure 2-1. Events are tallied only for $0.45 \leq q_1 \leq 0.55$.

event along edge (u, v) has equal probability to change $\mathcal{L}_u$ as $\mathcal{L}_v$. The expected number of Type 1 events is therefore equal to the expected number of Type 3 events, and we have $\mathbb{E}[E|i] = \mathbb{E}[G|i] = 1 - \phi_{K_0}(\beta_i)$. Finally, we compute the expected number of Type 2 events. By definition, for a Type 2 event to occur, the edge must no longer be attached to u. The expected number of such edges is $\mathbb{E}[K_0 + J_0 - G(K + J)|i]$. The probability that such an edge was removed by u by a rewiring event that did not deactivate the edge is σ_i. We obtain

$$\mathbb{E}[F|i] = \sigma_i \mathbb{E}[K_0 + J_0 - G(K + J)|i] \, .$$

An important prediction of this formalism is that Type 1 and Type 3 events, though they occur at the same rate, have different impacts on the active edge density. Since $\mathbb{E}[K|i] < \mathbb{E}[K_0|i]$ and $\mathbb{E}[J|i] > \mathbb{E}[J_0|i]$, we have

$$\mathbf{e}_i(\mathbf{c})_{\bar{\imath}i} = \frac{\mathbb{E}[K_0|i] - \mathbb{E}[J_0|i]}{2} > \frac{\mathbb{E}[K|i] - \mathbb{E}[J|i]}{2} = -g_i(\mathbf{q}, \mathbf{c})_{\bar{\imath}i} \, . \tag{2.8}$$

Equation (2.8) states that Type 1 events increase the active edge density more than

Term	Expression
Type 1 expected increment	$\mathbf{e}_i(\mathbf{c})_{01} = c_{\bar{i}i} - c_{ii} - 1$
Type 2 expected increment	$\mathbf{f}_i(\mathbf{c})_{01} = \frac{\mathbf{e}_0(\mathbf{c})_{01} + \mathbf{e}_1(\mathbf{c})_{01}}{2}$
Type 3 expected increment	$\mathbf{g}_i(\mathbf{c})_{01} = c_{\bar{i}i} + \varepsilon_i \frac{\beta_i}{1-\beta_i}\left(1 - \phi_{K_0}(\beta_i)\right)$
Type 1 expected count	$\mathbb{E}[E\|i] = 1 - \phi_{K_0}(\beta_i)$
Type 2 expected count	$\mathbb{E}[F\|i] = \sigma_i(c_{\bar{i}\bar{i}} + c_{\bar{i}i} - \mathbf{g}_i(\mathbf{c})_{01})$
Type 3 expected count	$\mathbb{E}[G\|i] = 1 - \phi_{K_0}(\beta_i)$

Table 2.1: Summary of the terms appearing in Equation (2.9). Only the 01 components (corresponding to active edges) are shown.

Type 3 events decrease it. This reflects a local asymmetry in the subcritical regime, between votes that increase the prevalence of a local minority opinion and votes that reduce it. The asymmetry is due to the intervening rewiring-steps, which tend to remove edges from the focal node u prior to a Type 3 event. Since Type 1 and Type 3 events occur at the same rate, our formalism predicts that voting events tend to increase the active edge-density when ρ is small. In Figure 2-2, we check this prediction by comparing the expressions in Equation (2.8) to the distribution of all impacts ΔM_{01} on the active edge count due to voting events. In the subcritical regime, the mean increment (black) is positive, reflecting the fact that Type 1 events (blue) outstrip Type 3 events (orange) in expected generation of active edges. As ρ grows, the separation-of-timescales assumption degrades, and the asymmetry between Type 1 and Type 3 events breaks down. For large ρ, Type 1 and Type 3 events have similar increments in expectation and the distribution of ΔM_{01} becomes symmetric.

Finally, we average over events of Types 1-3 to obtain the approximate expected increment in edge counts per voting event. It is given by the four-vector

$$\hat{\mathbf{v}}(\mathbf{q}, \mathbf{x}) = \frac{1}{2} \sum_{i \in \{0,1\}} \frac{\mathbb{E}[E|i]\mathbf{e}_i(\mathbf{c}) + \mathbb{E}[F|i]\mathbf{f}(\mathbf{c}) + \mathbb{E}[G|i]\mathbf{g}_i(\mathbf{q}, \mathbf{c})}{\mathbb{E}[E + F + G|i]} \,. \tag{2.9}$$

For convenience, we summarize the expressions appearing in Equation (2.9) in Table 2.1.

Combining Equation (2.1) with Equations (2.3), (2.4) and (2.9) and setting the

lefthand side equal to zero gives our Markovian approximation for the arch:

$$0 = \lambda \mathbf{w}(\mathbf{x}) + (1-\lambda)\alpha \mathbf{r}(\mathbf{q}) + (1-\lambda)(1-\alpha)\hat{\mathbf{v}}(\mathbf{q}, \mathbf{x}) \, . \tag{2.10}$$

This is a closed, deterministic equation in $\mathbf{x}$, derived under assumptions that are approximately correct near the critical point. Solving this equation yields $\hat{\mathbf{x}}$, the limit point of the approximate dynamics under Equation (2.10).[3] The approximation indicates the subcritical case when $\hat{\rho}(\mathbf{q}; \alpha, \lambda) = 2\hat{\mathbf{x}}_{01}(\mathbf{q}; \alpha, \lambda) \leq 0$, and the supercritical case otherwise. Solving

$$\alpha^*(\mathbf{q}, \lambda) = \max\{\alpha : \hat{\rho}(\mathbf{q}; \alpha, \lambda) = 0\} \tag{2.11}$$

then gives our approximation for the critical value α^* at which persistent disagreement emerges. We again emphasize that this solution is independent of both the time step t and the initialization of $\mathcal{G}$.

Figure 2-3 compares numerical solutions to Equation (2.11) simulation data for the complete range of $q_1 \in [0, 1]$. The accuracy of the approximation is strongest for $\mathbf{q} \approx \left(\frac{1}{2}, \frac{1}{2}\right)$ and on the rewire-to-random model variant. See also Figure 2-1(a) for comparisons of the solutions of Equation (2.11) to extant approximation schemes in the case $\mathbf{q} = \left(\frac{1}{2}, \frac{1}{2}\right)$.

Figure 2-3 highlights one of the qualitative differences between the rewire-to-random and rewire-to-same model variants. As discussed in Section 2.2, while α^* depends strongly on $\mathbf{q}$ in the rewire-to-random model variant, it is independent of $\mathbf{q}$ in the rewire-to-same variant. This behavior is reflected algebraically in Equations (2.7) and (5.7), which in turn govern the terms appearing in Equation (2.10). The quantities β, ε, and σ depend directly on $\mathbf{q}$ in the rewire-to-random model, regardless of the value of $\mathbf{x}$. However, in the rewire-to-same model, dependence on $\mathbf{q}$ emerges only when $\rho > 0$. This in turn implies that the phase transition is itself independent of $\mathbf{q}$, as is indeed observed in both the data and our approximation.

[3] In principle, Equation (2.10) may admit multiple limit points. Throughout our numerical numerical experiments, we have found the limit point to be unique.

Beyond the algebra, the localized approximation scheme we have developed gives to our knowledge the first mechanistic explanation of this difference in the phase transitions of the two models.[4] Consider the emergence of dissenting node u with opinion 1 on a component of majority opinion 0. In the rewire-to-random model, the fast local rewiring dynamics depend explicitly on $\mathbf{q}$, the global opinion densities. When q_1 is large, an edge rewired away from u is more likely to become inactive, resulting in fewer active edges in the neighborhood of u. This is in turn reflected by the term $g_1(\mathbf{q}, \mathbf{c})_{01} = \frac{1}{2}(\mathbb{E}[K|i] - \mathbb{E}[J|i])$ governing the impact of Type 3 events, whose magnitude enters into the calculation of the phase transition via Equations (2.10) and (2.11). In the rewire-to-same case, however, the fast rewiring does not explicitly depend on $\mathbf{q}$. An active edge attached to u that rewires becomes inactive with probability 1. As a result, there is no dependence of Type 3 events on $\mathbf{q}$, and the phase transition is independent of $\mathbf{q}$.

We now turn to the approximation of $\rho^*(\mathbf{q}; \alpha, \lambda)$, the equilibrium density of active edges in the supercritical regime. In this regime, the distinction between local minority and majority nodes progressively erodes, as does the validity of the timescale-separation assumption. One way to view this erosion is in terms of decay of the impact of Type 3 events, as discussed in Figure 2-2. As ρ increases, the impact of a single Type 3 event progressively diminishes due to re-randomization of the focal node's local neighborhood. We model this re-randomization via a simple interpolation to a mean-field approximation of the arch in the case $\alpha = 0$, which corresponds to a variant of the voter model without rewiring. We begin by deriving this approximation.

When $\alpha = \lambda = 0$, active edges enter and exit the system only through voting events. We have already written the mean-field approximation for the expected impact of a voting event in Equation (2.3). When only these events take place, the equilibrium

[4] The pair-approximation (PA) equations of [67] predict this difference but the mechanism therein is less clear to us.

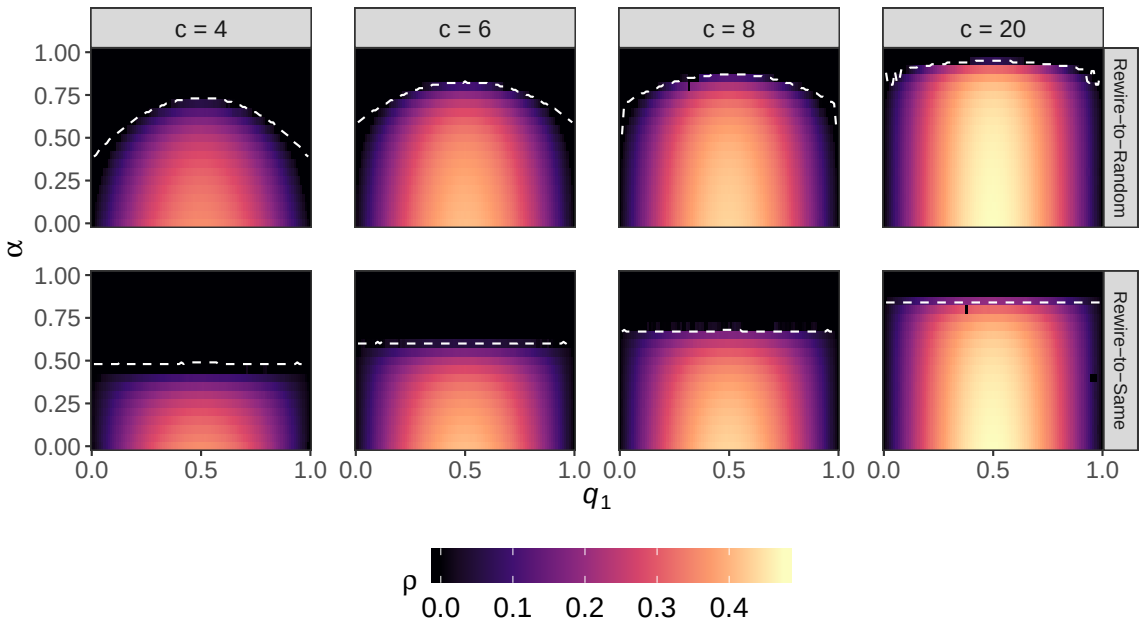


Figure 2-3: Approximation of the phase transition α^* for rewire-to-random and rewire-to-same systems for varying c and $\mathbf{q}$.
Color gives the equilibrium density of active edges. Dashed lines give solutions to Equation (2.11). Some numerical artifacts are visible in the rewire-to-random case for large c. Simulations were carried out under the conditions described in Figure 2-1.

condition is $e_i(\mathbf{c}) = 0$ for $i = 0, 1$. It suffices to solve the system

$$0 = 1 + c_{10} - c_{00}$$
$$0 = 1 + c_{01} - c_{11}$$

for $\mathbf{c}$ and subsequently for $\mathbf{x}$. We recall that $c_{ij} = cx_{ij}/q_i$ and that $2x_{01} = 1 - x_{00} - x_{11}$. Substituting these relations we obtain

$$\frac{2q_0q_1}{c}\begin{pmatrix}1\\1\end{pmatrix} + \mathbf{q} = \begin{bmatrix}1+q_1 & q_0\\ q_1 & 1+q_0\end{bmatrix}\begin{pmatrix}x_{00}\\x_{11}\end{pmatrix}.$$

The unique solution is

$$\begin{pmatrix}x^*_{00}\\x^*_{11}\end{pmatrix} = \frac{q_0q_1}{c}\mathbf{e} + \frac{1}{2}\begin{pmatrix}q_0(1+q_0-q_1)\\q_1(1+q_1-q_0)\end{pmatrix}.$$

We may then compute the mean-field approximation for the $\alpha = 0$ arch:

$$\hat{\rho}^*(\mathbf{q}) = 2x_{01} = 1 - x^*_{00} - x^*_{11} = 2q_0q_1\frac{c-1}{c}.$$

We note that this result is identical to that derived in [203] for a node-updating non-adaptive voter model.

We now introduce the interpolation function

$$s(\mathbf{q}, \mathbf{x}) = \frac{\hat{\rho}^*(\mathbf{q}) - \rho}{\hat{\rho}^*(\mathbf{q})} \tag{2.12}$$

to quantify the distance of the system state from the estimated $\alpha = 0$ arch. We then use this interpolation function to introduce decay in Type 3 events, replacing $\mathbf{g}(\mathbf{q}, \mathbf{c})$ in Equation (2.9) with $\tilde{\mathbf{g}}(\mathbf{q}, \mathbf{c}) = \mathbf{g}(\mathbf{q}, \mathbf{c})s(\mathbf{q}, \mathbf{x})$. The corresponding solution for $\hat{\mathbf{x}}$ yields the supercritical approximation of ρ.

Figure 2-4 shows the resulting approximations for the arch in both models, across a range of parameter regimes and both model variants. The arches for the rewire-

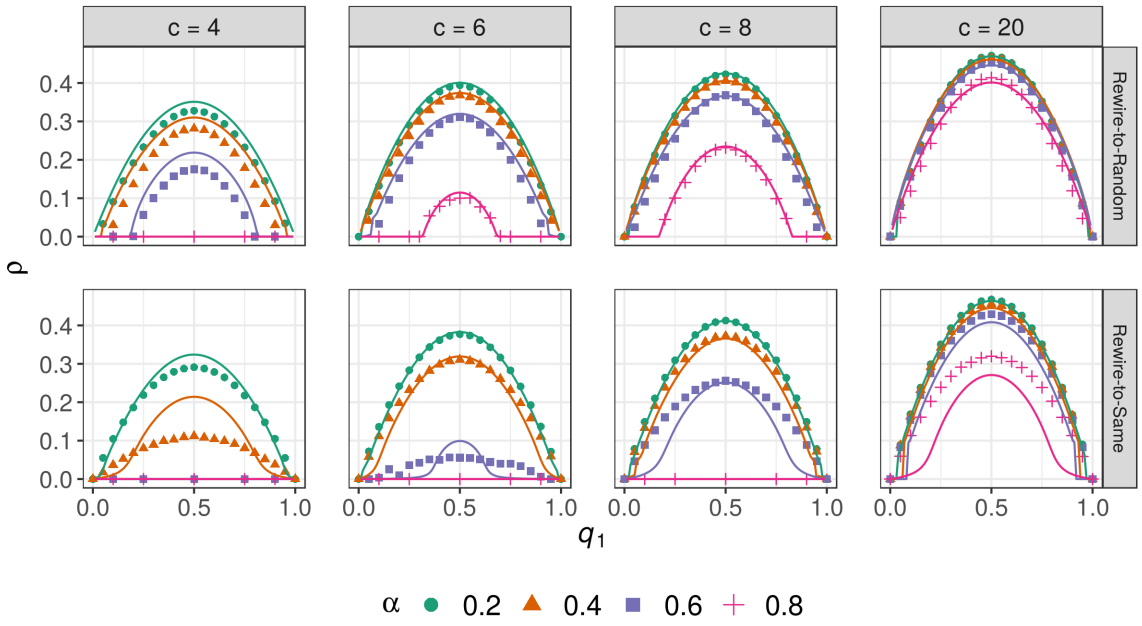


Figure 2-4: Approximations to the arch for varying α, $\mathbf{q}$, and c.
Points give averages over simulation runs on AVMs under the conditions described in Figure 2-1. Solid lines give the equilibrium value of $\hat{\rho}$ obtained by numerically solving for the fixed points of Equation (2.1) using the interpolation function given by Equation (2.12).

to-random model agree well with the data on both the support of the arch and the equilibrium active edge density. The rewire-to-same arches are somewhat less precise. The arches do correctly span the complete interval $[0, 1]$. The overall numerical agreement with the data is comparable to extant methods, but the parabolic shape of the arch is not completely reproduced — there is some warping near the base. The reason for this warping is not clear to us at present, and further investigation into this phenomenon may yield progress both theoretical and computational.

2.4 Discussion

The Markovian approximation technique we have developed offers predictions for the equilibrium active edge density ρ^* across the entirety of parameter space, and for varying opinion densities $\mathbf{q}$. Its accuracy in these tasks is generally comparable to that of the best extant methods. For example, Figure 2-1(a) shows that our Markovian

approximation is at least as accurate as AMEs [67] in predicting the $c = 4$ phase transition for the rewire-to-random model, and grows more accurate as c grows large. Our approximation is substantially more accurate than AMEs for the rewire-to-same phase transition, and only slightly less accurate than the compartmental approach of [35] for this model variant. Relatively few approximation schemes make predictions for the full arch, and it is therefore more difficult to make quantitative comparisons. The compartmental method of [63] approximates the equilibrium active edge density at $\mathbf{q} = \left(\frac{1}{2}, \frac{1}{2}\right)$ in the rewire-to-same variant more accurately than our method (Figure 2-1(b)), but does not make predictions for asymmetric opinion densities. AME predictions [67] recover the asymmetric phase transition and arches reasonably well in the rewire-to-random case, but are much less accurate for the rewire-to-same variant. Whereas the AME arches display warping in the rewire-to-random variant, our Markovian approximation displays warping in the rewire-to-same variant. In the $c = 4$ case shown in Figure 2-1(b), the present method offers overall accuracy in computing the arches similar to that of the AMEs, and improves as c grows large.

2.4.1 Computational Considerations

Solving eq. (2.10) requires finding the solution of a system of four coupled nonlinear equations, which may be done efficiently using a standard numerical solver. Notably, the dimensionality of the approximation is independent of the mean degree c. This contrasts to compartmental methods [67, 35, 189, 63], the dimension of which generally display quadratic or higher scaling in c. For example, AMEs comprise a system of $\Theta\left(k_{\max}^2\right)$ coupled differential equations, where $k_{\max}$ is the highest node degree expected to be encountered in simulation; in the $c = 4$ case, the authors of [67] used 272 such equations. This scaling makes the computation of approximations computationally prohibitive even for modest mean degree c. Similarly, the method of [35] for approximating the rewire-to-same phase transition requires a bisection search in α for which each function evaluation corresponds to finding the largest eigenvalue of a $(c-1)\times(c-1)$ matrix. The scaling is thus at least $O\left((\log\frac{1}{\epsilon})(c-1)^2\right)$, where ϵ is the desired approximation accuracy. Because our proposed method scales independently

of c, it can be used to approximate AVMs with arbitrarily large mean degrees.

2.4.2 Conclusions

Adaptive voter models offer a simple set of mechanisms that generate emergent opinion-based assortativity in complex networks. While the underlying rules are simple to state, the coevolving nature of the dynamics render these systems interesting and challenging to analyze. We have considered an ergodic adaptive voter model variant which enables a local perspective on fragmentation transitions and other model properties. The local perspective allows us to use the asymmetry of voting events to develop Markovian approximations based on the fast timescale dynamics around single nodes. The resulting approach is conceptually intuitive, computationally tractable, and predictively performant.

One of the most puzzling issues raised by our results is the difference between the accuracies of our approach for the rewire-to-random and rewire-to-same adaptive voter models. While we succeed in characterizing the rewire-to-random arch nearly exactly, the same methods produce poorer results for the rewire-to-same model. We conjecture that the rapid local sorting produced in rewire-to-same dynamics violates our mean-field assumption on Type 1 events, which would lead to approximation degradation. It would be interesting to extend our methodology to see whether refinements are possible that better characterize the rewire-to-same behavior and shed further light on the essential features governing the dramatic difference in the nature of the phase transitions in these two models.

It is also of interest to consider extensions and generalizations. The most natural extension is to the case of multiple opinion states and structured opinion spaces. Previous work on multi-opinion models has been restricted to either approximation of the various phase transitions [36] or empirical discussion of supercritical equilibrium behavior [186]. One reason for this is computational. The number of operations required to compute approximations under active-motif and AME approaches are exponential in the number $|\mathcal{X}|$ of opinion states, rendering both methods infeasible. In contrast, likely extensions of our local Markovian approximation methods scale as

$\Theta(|\mathcal{X}|^2)$, which would offer a significant reduction in computing time. If accuracy were preserved, such extensions would present the first scalable analytic methods for multi-opinion models. Other generalizations are also possible. While we developed our approximations for the specific case of the binary-state AVM, that development relies only on ergodicity, timescale-separation, and the mean-field assumption. We conjecture that these ingredients should be present in any adaptive model with homophilic dynamics in which rewiring steps involve uniform selection from an extensive subset of the graph, such as a subset sharing a given node state. An example of a more complex system in which these ingredients are present is the networked evolutionary prisoner's dilemma game of [123], in which nodes display richer strategic behavior in their opinion update and rewiring behavior. The existence of a phase transition driven by homophily may allow for the deployment of our novel methods in such cases as well.

Code

Documented code for running simulations and computing the approximations described in this paper may be found at `https://github.com/PhilChodrow/AVM`.

Chapter 3

The Structure of Spatial Segregation

The ethnic composition of America is shifting: the white population, which composed nearly 85% of America in 1965, will account for less than 50% by 2050 [164]. Ethnic diversification, however, has been accompanied by evolving forms of racial segregation. White-Hispanic and white-Asian segregation are on the rise [81], even as white-black segregation has declined [127, 73, 72]. Understanding and intervening in these shifts in segregation is a major priority for modern American cities. In addition to the moral issues associated with unequal access to education, jobs, and public resources, racial segregation is economically destructive to entire cities. A recent study [2] estimated that stark white-black segregation in Chicago has annual costs of $3.6B in income among black residents; over 80,000 college graduates; and over one hundred homicides.

Computational tools for the detailed study of segregation are, however, limited. Most quantitative methods have centered around the quantification of *degrees* of segregation along different conceptual axes. The growing availability of high-resolution demographic data has generated interest in a richer class of problems and methods among researchers studying cities. In this essay, we will consider two main classes of problem:

Quantifying Spatial Scale: How do cities differ in the *scale* on which they are segregated? Is there a single spatial scale of demographic separation–such as in Figures 3-1a and 3-1b– or multiple scales – such as in Figures 3-1c and 3-

1d? How can we quantify average spatial scale and make comparisons between cities?

Learning Spatial Structure: How can we define algorithms that automatically learn spatial structure in segregated cities? How can we compare spatial structure between cities? How can we carry out dimension-reduction that preserves coarse-grained spatial structure for agent-based modeling or inferential statistics?

These and similar questions have surfaced in recent studies ranging from sociology [170, 122] to physics [129, 28]. Reliable methods for answering these questions would support physics-style studies of studies of urban morphology and information-scaling; studies of the relationship of ethnic geography to patterns of disparity or violence [126]; and more traditional sociological studies in which spatial regions must be non-arbitrarily defined for inferential analysis. However, a satisfactory methodological framework for addressing these questions remains elusive.

This chapter is structured as follows. In Section 4.1, we survey the history of the methodology of measuring segregation, focusing on the literature coming from quantitative sociology. We highlight both the progressive advancement of this methodology, as well as some of its outstanding limitations. In Section 3.2, we present the mathematical foundations of our methodology. Our core technology is the algebra of Bregman divergences and their well-known connections to Riemannian geometry. We show that this technology allows us to subsume most state-of-the-art smoothing-based segregation measures as special cases corresponding to particular smoothers and choices of divergence. An important virtue of this abstract approach is that the methods we formulate are equally applicable to both unordered variables (such as ethnicity) and ordered variables (such as income or educational attainment). We then use this geometry to construct the local information density, a measure of the average spatial scale of demographic variation. The local information density may be viewed as a localization of the mutual information, and we prove a geometric theorem relating the mutual information to the metric tensor that governs the computation of distances

in information space. Importantly, this theorem both gives intuition on the nature of the local information and allows us to approximate it without recourse to ill-defined limits in discrete data space. We next develop multiscalar computational methods for identifying boundaries in demographic data. We formulate two algorithms, which can be used in tandem, for detecting these boundaries. The agglomerative method is of particular interest, as it exploits the decomposition properties of Bregman informations to detect quantified, hierarchically-nested levels of segregation.

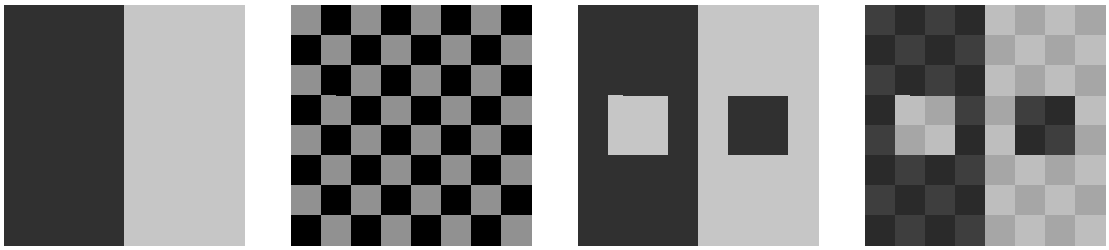
In Section 3.3, we deploy our methods on blockgroup-level data on ethnoracial demographics provided by the American Community Survey [1]. We choose 56 large American cities for certain comparative purposes, though we focus on Detroit, Chicago, and Philadelphia for more in-depth analysis and illustrations. We conclude in Section 5.6 with a discussion of several directions of future work.

A much-condensed version of this chapter was published as "Structure and Information in Spatial Segregation" in *The Proceedings of the National Academy of Sciences* [47].

3.1 Segregation Measures: From Aspatial to Multiscalar

To motivate the need for unified, novel methodology, we now take a brief survey of recent work in the study of the spatial demographic structure of cities.

Historically, methodological work in segregation studies has focused on quantifying the *degree* of segregation in a given city or system. This work has given rise to a proliferation of indices: scalars that can be computed from data to give an overall measure of the extent to which a given region is segregated. Prior to the widespread availability of computational resources, most segregation indices were aspatial. A typical example, still studied today [80, 171, 175, 28], is the mutual information between spatial locations and ethnic groups. Let $\mathcal{Y}$ be finite alphabet of ethnic groups, and let $\mathcal{X}$ be a finite set of spatial locations, such as Census tracts. We



(a) (b) (c) (d)

Figure 3-1: Illustration of multiscale segregation patterns in four toy cities. 3-1a: Segregation on a single, large spatial scale. 3-1b: Segregation on a single, smaller spatial scale. 3-1c and 3-1d: Segregation on two and three hierarchical spatial scales.

view the population of the city as a joint distribution $p_{X,Y}$ over $\mathcal{X} \times \mathcal{Y}$, in which $p_{X,Y}(x, y)$ is the probability that an individual sampled uniformly at random from the city lives at location x and is a member of ethnic group y. Then, the mutual information between locations X and ethnic groups Y is given by

$$I(X,Y) \triangleq \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} .$$

Intuitively, $I(X,Y)$ takes its maximum value in a diverse city when each spatial location is monoracial, and its minimum when each location is identical in demographic composition.

Importantly, the mutual information so defined is aspatial, in the sense that it is invariant under permutations of $\mathcal{X}$. It therefore does not distinguish between the two "checkerboard" cities 3-1a and 3-1b shown in Figure 3-1, since the latter is a spatial permutation of the former. As computational resources grew and spatial analysis software developed, researchers began to formulate measures that would distinguish between cities 3-1a and 3-1b in operationally-meaningful ways. Their recent efforts may be divided into two principle approaches. The *geostatistical* approach, championed primarily by David Wong in a series of related papers, [213, 214, 158], attempts to derive spatial measures of segregation from geographic and geometric reasoning, typically by studying the spatial distributions $p_{X|Y=y}$ of a given group y over a spa-

tial region. The distributions for each group may then be compared; segregation between two groups is high if their spatial distributions differ under an appropriately formulated distance metric. These approaches have the useful property of generating easily-understood visualizations in addition to indices of segregation, thereby facilitating analysts in exploring not just the magnitude but also the structure of spatial segregation. However, geostatistical methods also possess significant limitations. Because such approaches use only information contained in the conditional distributions $p(X|Y = y)$, the indices so computed enjoy no privileged status in probability theory, making their properties difficult to discern. Furthermore, as the authors of [158] acknowledge, it is unclear how to apply geostatistical methods to multigroup segregation; prospects for the geostatistical study of segregation based on ordered variables such as income are similarly unclear.

The second approach to studying the spatial structure of segregation "spatializes" aspatial measures like (3.2) via the use of a geographic smoothing kernel [173, 176, 108]. Such *smoothing-based* approaches thus consist of two methodological ingredients. The first is an aspatial index that acts on the joint distribution $p_{X,Y}$ to produce a scalar quantifying the degree of segregation. Appropriate aspatial measures include the mutual information (3.2) [171, 175, 80, 200], the local quadratic exposure [171], and the Neighborhood Sorting Index [108]. The second methodological ingredient is a spatial smoothing function Φ that acts on the joint distribution $p_{X,Y}$ to produce a modified joint distribution $\tilde{p}_{X,Y} = \Phi(p_{X,Y})$. This computation is generally performed via a smoothing kernel ϕ:[1]

$$\tilde{p}(x_0, y) = \int \phi(x_0, x) p(x, y)\, dx\ .$$

The analyst is required to specify the functional form of the kernel ϕ, thereby determining an appropriate notion of "closeness" in geographic space. Standard choices for the smoothing kernel ϕ include the Gaussian radial basis function used frequently in statistical learning; the truncated quartic smoother [176]; and graph-based measures

[1] In analytic practice, we have access to only a discrete number of points in $\mathcal{X}$, in which case the integral above is replaced by the corresponding sum.

that compute distances on the adjacency network of spatial tracts [108]. Many other kernels are possible; for example, recent advances in the study of the properties of human mobility [190] may suggest the use of alternative, physically-motivated kernels based on the statistics of Levy flights.

Aspatial measures have strong mathematical properties that make them attractive for practitioners [173, 175, 80], and these properties are inherited by their spatially smoothed counterparts. The smoothing approach is thus extremely promising for the measurement of average segregation experienced by individuals. Smoothing-based measures, however, are less appropriate for the study of the questions of structure and scale we posed above. Indeed, while smoothing-based measures have been used to study spatial scale [122, 170], there are substantial mathematical limitations to such approaches. Such studies proceed by computing smoothing-based measures under varying bandwidths ℓ for some fixed functional form of the smoothing kernel ϕ_ℓ. As we show below, most of the measures used in such studies may be viewed as forms of the mutual information $I(X, Y)$ between space and demographics. The application of the smoothing kernel amounts to a form of data pre-processing, and the data processing inequality implies that larger bandwidths will always tend to decrease segregation measures:

$$\ell_1 \leq \ell_2 \implies I \circ \phi_{\ell_1} \geq I \circ \phi_{\ell_2} \,. \tag{3.1}$$

While this property may be workable for the purposes of studying degrees of segregation at varying scales, it is clearly a limitation for dimension-reduction. When determining the number k of clusters appropriate for summarising the structure of Figure 3-1a, for example, we would expect a structure measure to be high at the largest spatial scale corresponding to $k = 2$ clusters. Smoothing-based aspatial measures, however, cannot display this behavior.

Spatial indices are also restrictive one when desires to study segregation on multiple spatial scales. For example, Figure 3-1c depicts a toy city with two, hierarchically-nested scales of segregation, and Figure 3-1d a toy city with three. Much recent work has therefore focused on the development of *multiscalar* methodology. Multiscalar

analysis emphasizes that different features of urban segregation are visible only at certain scales of analysis, where a "scale" may refer to a characteristic geographic length [170, 172, 181, 122]; a number of neighbors [157, 53, 156, 98]; or a set of aggregate spatial units such as Census tracts or designated places [77, 125]. Multiscalar approaches have traditionally been pursued through the study of *segregation profiles*. These are curves that plot the value of segregation index – usually the Information Theory Index [200] – as a function of a geographic smoothing bandwidth or number of nearest neighbors. These profiles efficiently represent how the degree of segregation depends on the size of the "local neighborhoods of individuals, and, in the case of [98], can also characterize the dependence of these values on the selection of the entire region of analysis. These methods, however, also share a characteristic limitation. Profile curves view scale as a global property – at each point on the curve, a single scale is used for the entire analytical region. Inspection of segregation in Detroit (Figure 3-2) suggests the coexistence of multiple scales of separation in different areas of the city. Because profile methods use a global scale, at each point on the profile curve, some of the local features of segregation in Detroit are necessarily lost. A recent paper [181] makes progress against the global scale limitation by constructing egocentric profile curves and studying their properties using clustering and inferential methods. This approach illuminates interesting spatial patterns of demographic difference, but decouples those egocentric profiles from overall segregation measures. Profile methods are also limited in their ability to characterize how much of overall segregation lies at any particular geographic scale. While measures like the macro-micro segregation ratio and net micro-segregation of [122] can suggest scalar decompositions of this type, they do not share with the explicit decomposition methods below the strong mathematical properties necessary to make such analysis precise.

Another approach to multiscale measurement explicitly *decomposes* overall segregation into terms reflecting contributions at different scales using a mathematical property of the Information Theory Index. Unlike profile methods, the decomposition approach does not assume a global geographic or population scale. Figure 3-2

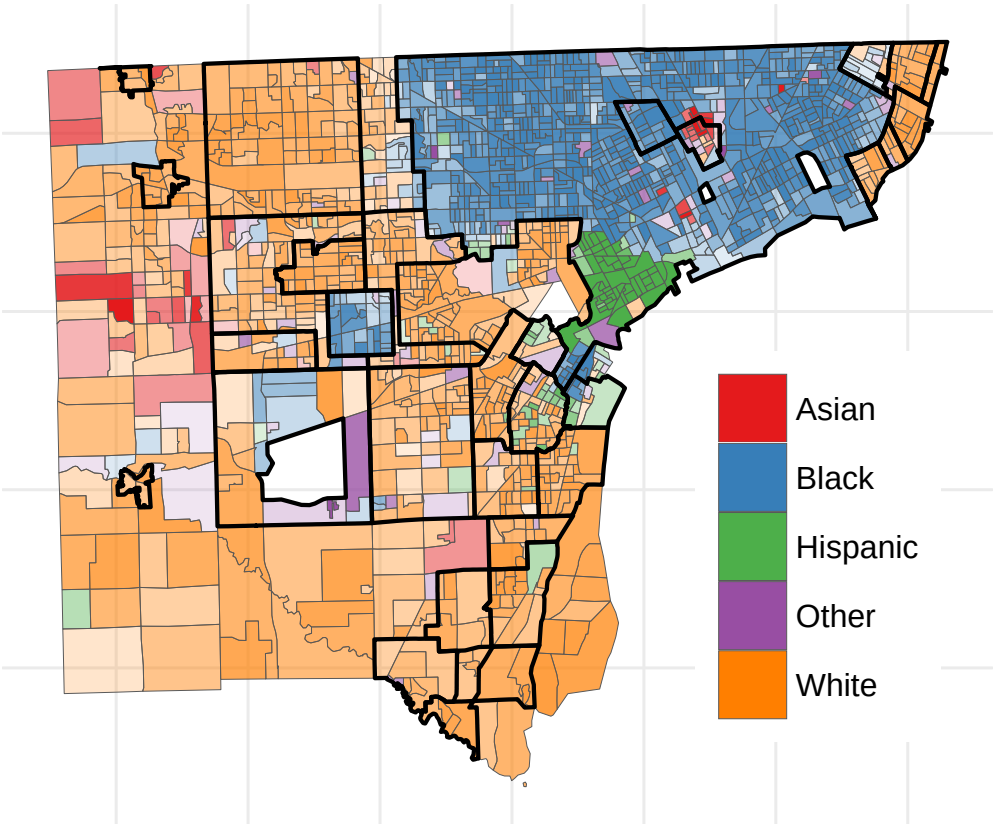


Figure 3-2: Census block groups in Wayne County, Michigan, including the city of Detroit.
Black boundaries delimit Census-designated places within the county. Each block-group is colored according to the group that is most concentrated in that blockgroup relative to the city-wide average, and the shade of each blockgroup corresponds to the degree of concentration.

illustrates this flexibility in Wayne County, with black outlines delimiting Census-designated places used as an intermediate scale. The designated places span a wide range of geographic and population scales, and many of their boundaries correspond to lines of demographic separation. A typical analysis proceeds to decompose overall segregation into between-place and within-place components, with the latter term reflecting segregation on a level "below" the scale of places. This approach to analysis is intuitive and addresses the problem of distinguishing segregation on multiple scales provided the availability of geographic units at each scale. Unfortunately, this dependence on the availability of appropriate geographic units is highly limiting. While [77] argues that places correspond to meaningful communities and administrative units, the use of places may also obscure important, large-scale features. Using the Information Theory Index, the place decomposition uses 30 distinct spatial units to capture just 48% of segregation in the between-place term, suggesting that more than half of overall segregation in Detroit lies below this scale. This suggestion may mislead. As we will see, just seven, different spatial units suffice to capture 71% of overall segregation. The use of fixed spatial units such as places may thus substantially understate the scale on which segregation processes operate.

This limitation is closely connected to the long-standing Modifiable Areal Unit Problem (MAUP) [155]. MAUP articulates the problem that many standard geographic units – such as grids, political demarcations, or administrative boundaries – may divide the data into categories that are arbitrary with respect to the phenomena under investigation. The checkerboard again provides a useful illustration: in Figure 3-1b, if we were to divide the toy city into 16 squares of size 2×2, each would contain the same demographic distribution and standard methods would conclude that no segregation exists. When studying spatial segregation using areal units, it is therefore necessary that the areal units either (a) be in some sense well-matched to the phenomenon studied or (b) at sufficiently high resolution that the researcher can strategically aggregate them.

In the next sections, we develop a unified mathematical framework for such study. This framework combines the strong conceptual core of smoothing-based measures

with tools from contemporary statistics and machine learning. Information theory [185, 59, 60] forms the foundation of our methodology. In addition to playing a major role in many aspatial and smoothing-based approaches discussed above [80, 175, 176, 171, 173], information theory has also been deployed in a number of other applications in the study of cities. One venerable strand of development has used entropy maximization and related concepts to study overall spatial population distributions [20, 19, 21, 22, 212]. More recent work deploys information measures [28] and elementary inference techniques [129] to study how different population groups are arranged in space.

3.2 Local and Global Scales

3.2.1 Compositional Triples

We begin by specifying the core mathematical environment in which we will operate. At suitable level of abstraction, a spatially-distributed population with varying attributes may be modeled with a geographic region, a population density function over that region, and a map that assigns to each location an empirical probability distribution over demographic attributes. Fix n and m, and let λ be the Lebesgue measure on $\mathbb{R}^n$.

Definition 2. *A **compositional triple** $\gamma = (R, \mu, \alpha)$ consists of:*

1. *A geographic region $R \subset \mathbb{R}^n$, which is compact and of nonzero measure under λ.*

2. *A population measure μ, which is a probability measure on R and is absolutely continuous with respect to λ.*

3. *An attribute map $\alpha : R \to \mathcal{P}$, which is continuously differentiable. The probability simplex $\mathcal{P}$ is defined as*

$$\mathcal{P} \triangleq \left\{ q \in \mathbb{R}^m \mid \sum_i q_i = 1 \ , \ q_i \geq 0 \ \forall i \right\} .$$

Compositional space Γ *is the set of all compositional triples.*

From the hypothesis of absolute continuity, the Radon-Nikodym Theorem implies that μ possesses a Radon-Nikodym derivative (density function) $\rho \triangleq \frac{d\mu}{d\lambda}$.

Definition 3. *An* ***index*** ψ *is a map* $\psi : \Gamma \to \mathbb{R}$ *that assigns to a given compositional triple* γ *a scalar* $\psi(\gamma)$.

The formalism of compositional triples allows a neat classification of many existing measures of population structure according to which elements of γ are used in the computation of $\psi(\gamma)$. Measures of population dispersion, such as the spatial entropy of [20, 212], use only the geographic region R and the population measure μ; that is, $\psi(\gamma) = \psi(R, \mu)$. Aspatial segregation measures employ the induced population measure $\mu \circ \alpha^{-1}$ and the image $M = \alpha(R)$, but neither R nor α directly, allowing us to write $\psi(\gamma) = \psi(\mu \circ \alpha^{-1}, \alpha(R))$. The indices of [171, 169] fall into this category. Smoothing-based indices have the same structure as aspatial measures, but α with a spatial smoother ϕ, yielding $\psi(\gamma) = \psi(\mu \circ \phi^{-1} \circ \alpha^{-1}, (\alpha \circ \phi)(R))$. The indices of [173, 176, 108] measuring spatial segregation in race and income are such smoothing-based indices. Apparently, there is information being left on the table here: none of these classes of measures deal directly with the the structure of the map α. That α is not used directly in any of these measures is somewhat surprising, given that α is the most complete representation of the relationship between space and demographics. In broad strokes, our strategy is to probe the structure of spatial segregation by constructing indices and algorithms that exploit this previously-underutilized structure.

A note on notation: our setup is somewhat unconventional in the context of the segregation literature, where it is more common to consider a joint probability measure $\mathbb{P}_{X,Y}$ between the variable X encoding spatial location and the variable Y encoding demographic group identity. The value of $\mathbb{P}_{X,Y}(B, i)$ gives the probability that an individual selected uniformly at random from the city lives in the set B and belongs to demographic group i. This notation can be recovered from our formalism as

$$\mathbb{P}_{X,Y}(B, i) = \mu(B) \int_B \alpha_i \, d\mu \, .$$

Since Y is a discrete random variable and μ is absolutely continuous with respect to λ (by hypothesis), we obtain a Radon-Nikodym derivative $p_{X,Y}$ of $\mathbb{P}_{X,Y}$.

3.2.2 Bregman Divergences on $\mathcal{P}$

Because α is continuously differentiable, its image $M = \alpha(R)$ has the structure of a parameterized differentiable submanifold of $\mathcal{P}$ [142]. The manifold structure implies that we can take derivatives, measure distances, and, more generally, perform geometrical calculations on M. Differential geometry is the study of such techniques, and its application to submanifolds of $\mathcal{P}$ is known as *information geometry* [8], a relatively young field of mathematics that has made rich contributions to statistical methodology. Our fundamental program is to use elementary information-geometric techniques to study the structure of a compositional triple $\gamma = (R, \mu, \alpha)$.

The use of geometrical techniques in this context requires that we specify an appropriate distance function on $\mathcal{P}$. In our chosen context, the choice of such a function amounts to making a systematic judgment about when two points $p, q \in \mathcal{P}$ are "similar" or "different." This choice must be guided by the application domain. In this section, we motivate and introduce Bregman divergences on $\mathcal{P}$, a flexible class of distance-like functions that enable the comparison of elements of $\mathcal{P}$, and which suffice to induce the required metric structure. To motivate the introduction of Bregman divergences, consider the following two examples.

For categorical variables such as racial groups in which the group labels have no particular ordering, a standard comparison function used in segregation studies is the Kullback-Leibler divergence [175, 200]:

$$d_{kl}(p, q) \triangleq \sum_j p_j \log \frac{p_j}{q_j} .$$

The KL divergence is usefully interpretable as the estimation loss associated with believing q when the true "state of the world" is given by p [60]. The KL divergence is not appropriate in all contexts. In studying income disparity, in which demographic categories have ordered structure (e.g. {Poor, Middle-Class, Rich}), however, the

KL divergence is less appropriate. Because the KL divergence is invariant under permutations of group labels, its use in the study of ordered variables loses important information. For example, consider three sample demographic distributions:

$$p = \left(\frac{1}{2}, \frac{1}{2}, 0\right), \quad q = \left(\frac{1}{2}, 0, \frac{1}{2}\right), \quad r = \left(0, \frac{1}{2}, \frac{1}{2}\right).$$

If these are viewed as distributions over economic class, an appropriate comparison function for such an analysis should treat q as more similar to p than r to p; in other words, we should have $d(p,q) < d(p,r)$. The KL divergence is apparently ill-suited for this analysis, and gives $d(p,q) = d(p,r) = \infty$. A more appropriate comparison function is the cumulative Euclidean metric, given by

$$d_{ce}(p,q) = (p-q)^T C (p-q),$$

where C is the matrix

$$C = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & \cdots & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & \cdots & m \end{bmatrix}.$$

The cumulative Euclidean metric d_{ce} derives its name from the fact that if $c(p)$ is the vector of cumulative sums of p (i.e. $c(p)_k = \sum_{j \le k} p_i$), then $d_{ce}(p,q) = \|c(p) - c(q)\|^2$. In our example above, we may calculate

$$d_{ce}(p,q) = \frac{1}{4}, \quad d_{ce}(p,r) = \frac{1}{2},$$

confirming that d_{ce} reflects our intuitive judgment that q is more similar to p than is r.

Both d_{kl} and d_{ce} are members of a general class of comparison functions called *Bregman divergences* [40]. Individual Bregman divergences on $\mathcal{P}$ are characterized by a strictly convex, C^1 function $f : \mathcal{P} \to \mathbb{R}$. The divergence induced by f is then the

function $d_f : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ given by

$$d_f(p, q) = f(p) - f(q) - Df_q(p - q) \ ,$$

where Df_q is the (matrix) derivative of f at q. The KL divergence is induced by the negative entropy of p: $f(p) = \sum_i p_i \log p_i = -H(p)$. The cumulative Euclidean metric is induced by

$$f(p) = \frac{1}{2} p^T C p \ ,$$

and may be viewed as a member of the more general class of Mahalanobis distances [**?**]. The squared Euclidean distance is also a member of this class, and is induced by $f(p) = \frac{1}{2} \|p\|^2$. Bregman divergences are not themselves appropriate as distance measures for geometric purposes; however, as we will discuss in Subsection 3.2.4, the choice of f induces a unique Riemannian metric on $\mathcal{P}$ that will provide the appropriate notion of distance. Before this discussion, we pause to emphasize the role of Bregman divergences in existing measures of segregation and inequality.

Bregman divergences may be used to construct generalized versions of the entropy, conditional entropy, and mutual information of classical (Shannon) information theory. These constructions find frequent use in the measurement of segregation, and so we briefly develop them here. We will adopt the notation and terminology of Shannon theory, in which the divergence generating function f is the negative entropy. However, we emphasize that the validity of these methods do not depend on the functional form of f.

Let $\bar{q} \in \mathcal{P}$ be the mean demographic vector given by $\bar{q} = \int \alpha \, d\mu$.

Definition 4. *The* entropy *of demographic groups Y is* $H(Y) \triangleq -f(\bar{q})$.

Definition 5. *The* conditional entropy *of Y given X is* $H(Y|X) \triangleq -\int (f \circ \alpha) \, d\mu$.

Definition 6. *The* mutual information *between spatial location X and demographic groups Y is*

$$I(X, Y) \triangleq \int d_f(\alpha(x), \bar{q}) \, d\mu(x) \ . \tag{3.2}$$

Theorem 1. *The entropy, conditional entropy, and mutual information satisfy* $I(X,Y) = H(Y) - H(Y|X)$.

Proof. This fact is well-known in Shannon information theory; we show that it is independent of the choice of f. Starting with the righthand side,

$$\begin{aligned} H(Y) - H(Y|X) &= -f(\bar{q}) + \int (f \circ \alpha)(x)\, d\mu(x) \\ &= -f(\bar{q}) + \int \left[f(\bar{q}) + Df_{\bar{q}}(\alpha(x) - \bar{q}) + d_f(\alpha(x), \bar{q}) \right] d\mu(x) \\ &= \int Df_{\bar{q}}(\alpha(x) - \bar{q})\, d\mu(x) + \int d_f(\alpha(x), \bar{q})\, d\mu(x)\,. \end{aligned}$$

Since $\bar{q} = \int \alpha \, d\mu$, the first integral vanishes and the remaining second integral is simply $I(X,Y)$. □

Theorem 1 has an attractive interpretation in the context of segregation. The entropy $H(Y)$ may be viewed as a measure of global diversity; the mutual information $I(X,Y)$ as a measure of global unevenness [171, 80, 175, 200]; and the conditional entropy $H(Y|X)$ as a measure of local exposure. Theorem 1 then reads:

$$\text{Global Diversity} = \text{Local Exposure} + \text{Global Unevenness} \tag{3.3}$$

Intuitively diversity must reside either in "within neighborhoods" (first term) or "between neighborhoods" (second term). Because this relationship holds for general f, it holds for both unordered and ordered demographic variables according to the choice of the divergence function f.

3.2.3 Aspatial Measures of Segregation

We have intentionally developed these measures in a highly abstract setting to emphasize their generality. Indeed, most widely-accepted smoothing-based measures of inequality and segregation, in both categorical and ordinal variables, may be cast in terms of the above measures for appropriate choices of f and smoothing kernel ϕ. In Table 3.1, we organize the most-commonly used and cited smoothing-based mea-

Index	Functional Form	Divergence Generator	Smoother ϕ
Spatial Exposure [173]	$1 + H(Y\|X)$	$\|p\|^2$	General
Divergence [176, 175]	$I(X,Y)$	$\sum_i p_i \log p_i$	Quartic
Information Theory [200, 173]	$\frac{I(X,Y)}{H(Y)}$	$\sum_i p_i \log p_i$	General
Ordinal Information Theory [169]	$\frac{I(X,Y)}{H(Y)}$	$\sum_i c_i \log c_i$	General
Ordinal Variation Ratio [169]	$\frac{I(X,Y)}{H(Y)}$	$-4\sum_j c_j(1-c_j)$	General
Ordinal Square Root [169]	$\frac{I(X,Y)}{H(Y)}$	$-2\sum_j \sqrt{c_j(1-c_j)}$	General
Neighborhood Sorting [108]	$\sqrt{\frac{\tilde{I}(N,Y)}{I(X,Y)}}$	$(\sum_i p_i y_i)^2$	Ego-Network

Table 3.1: Taxonomy of smoothing-based segregation measures according to functional form, Bregman divergence generator f, and smoothing kernel ϕ.
In the three ordinal measures of [169], $c_i = \sum_{k=1}^{i} p_i$. The Neighborhood Sorting Index is designed for ordinal variables like income, in which y_i is an income level. In this measure, $\tilde{I}(N,Y)$ is the the mutual information between smoothed neighborhoods and demographics.

sures of segregation according to their information-theoretic structure. As Table 3.1 shows, the analyst need generally make just a few, well-defined choices in order to select a measure appropriate for their application. First is the use of an appropriate divergence-generator f, which encodes the domain-appropriate concept of diversity. Second is a spatial smoother ϕ, which encodes an operational notion of spatial proximity. Then, the sources above agree that, across a wide range of applications and contexts, the correct measure to compute is the mutual information $I(X,Y)$, potentially normalized by either global entropy (diversity) $H(Y)$ or a non-smoothed mutual information for the Neighborhood Sorting Index.

The theorem below proves the correctness of the entries in this table.

Theorem 2. *The following functions $f : \mathcal{P} \to \mathbb{R}$ are strictly convex on their domain and continuously differentiable on* int $\mathcal{P}$.

Euclidean Norm: $f_1(p) = \|p\|^2$.

Negative Entropy: $f_2(p) = \sum_i p_i \log p_i$.

Cumulative Entropy: $f_3(p) = (g_3 \circ \sigma)(p)$, *where* $g_3(c) = \sum_i c_i \log c_i$.

Cumulative Variance: $f_4(p) = (g_4 \circ \sigma)(p)$, *where* $g_4(c) = -4\sum_j c_j(1-c_j)$.

Cumulative Root Variance: $f_5(p) = (g_5 \circ \sigma)(p)$, *where* $g_5(c) = -2\sum_j \sqrt{c_j(1-c_j)}$.

Square Mean: *Assume additionally that the alphabet* $\mathcal{Y}$ *is an alphabet of integers, and let* $f_6(p) = (\sum_i p_i y_i)^2$.

The function $\sigma : \mathcal{P} \to \mathbb{R}^n_+$ *is the cumulative summation map given by* $\sigma(p)_k = \sum_{j=1}^k p_j$.

Proof. Continuous differentiability of each function is clear, since σ is a linear map. Strict convexity of f_1 and f_2 are well-known, so it remains to prove strict convexity of f_3 through f_6.

Cumulative Entropy: The map σ is linear, and f_3 is therefore strictly convex iff g_3 is strictly convex on $\mathcal{R}^n_+$. We note that

$$\frac{\partial^2 g_3(c)}{\partial c_i \partial c_j} = \begin{cases} \frac{1}{c_i} & \text{i} = \text{j} \\ 0 & \text{otherwise} . \end{cases}$$

We therefore have that $\mathcal{H}g_3(c)$ is diagonal and consists of positive entries everywhere in $\mathbb{R}^n_+$, and is therefore positive-definite as required.

Cumulative Variance: Convexity is unaffected by linear terms. Since $g_4(c) - 4e^T c = 4\|c\|^2$ is strictly convex on $\mathbb{R}^n_+$, so is g_4.

Cumulative Root Variance : We have

$$\frac{\partial^2 g_5(c)}{\partial c_i \partial c_j} = \begin{cases} \frac{1}{2(c_i(1-c_i))^{3/2}} , & i = j \\ 0 \quad \text{otherwise} \end{cases} .$$

As with $g_5(c)$, $\mathcal{H}g_5$ is therefore diagonal and consists of positive entries, and is therefore positive-definite.

Square Mean : The map $y \mapsto \sum_i p_i y_i$ is linear, and the squaring operation is strictly convex.

□

Corrolary 1. *The spatial exposure index between group m and group n is defined by [173] as*

$$P_{mn} = \int p(m|x)p(n|x)\, d\mathbb{P}_X(x).$$

The spatial exposure index satisfies

$$\sum_{m \neq n} P_{mn} = 1 + H(Y|X)\ ,$$

where the divergence generating function is the Euclidean norm $f_1(p) = \|p\|^2$.

Corrolary 2. *The Divergence Index D of [175, 176] satisfies*

$$D = I(X,Y)\ ,$$

where the divergence generating function is the entropy $f_2(p) = \sum_i p_i \log p_i$.

Corrolary 3. *The Information Theory Index $\tilde{H}$ of [200, 173, 169] satisfies*

$$\tilde{H} = \frac{I(X,Y)}{H(Y)}\ ,$$

where the divergence generating function is the entropy $f_2(p) = \sum_i p_i \log p_i$.

Proof. Equation (7) of [173], for example, defines $\tilde{H}$ as

$$\tilde{H} = \frac{H(Y) - H(Y|X)}{H(Y)}\ ,$$

from which the result follows by the standard identity $I(X,Y) = H(Y) - H(Y|X)$. □

Corrolary 4. *The Ordinal Information Theory Index H^O, Ordinal Variation Ratio RO, and Ordinal Square Root Index S^O of [169] are all of the form*

$$\frac{I(X,Y)}{H(Y)}\ ,$$

where the divergence generating functions are f_3, f_4, and f_5, respectively.

Corrolary 5. *Let N be a random variable with $H(N|X) = 0$; that is, N is completely determined by X. The Generalized Neighborhood Sorting Index G of [108] satisfies*

$$G = \sqrt{\frac{\tilde{I}(N,Y)}{I(X,Y)}},$$

where the divergence generating function is f_6, and where $\tilde{I}(N,Y)$ is computed with respect to the spatially smoothed distribution $\Phi(p)$, where Φ is the uniform ego-network smoother of order n.

Proof. Since we have already proved the convexity and continuous differentiability of f_6, the only task required is to cast Equation (4) of [108] in the claimed form. The numerator of this equation may be written

$$\int d_f(\tilde{p}(\cdot|n), \bar{p})\, d\mathbb{P}_N(n),$$

and the denominator may be written

$$\int d_f(\tilde{p}(\cdot|x), \bar{p})\, d\mathbb{P}_X(x),$$

as required. □

3.2.4 The Metric Tensor and Local Information Density

Having developed the role of Bregman divergences in structuring common smoothing-based measures, we now proceed to construct explicitly spatial methods through the geometry of Bregman divergences. Our first task will be to construct a measure of *local scale*. This measure will be large when a pattern of spatial segregation is more fine-grained. In the language of checkerboards, this measure will be able to distinguish between Figure 3-1a and Figure 3-1b. Operationally, what we will do is compute the mutual information over very small neighborhoods, and prove a theorem that allows us to approximate the small-neighborhood limit.

The key fact we exploit is that each Bregman divergence is closely related to a

Riemannian metric – a quadratic form on $\mathcal{P}$ that provides an inner product. We start by noting that Bregman divergences are locally quadratic, in the sense that

$$d_f(q+\delta, q) = \frac{1}{2}\mathcal{H}f_q(\delta,\delta) + o\left(\|\delta\|^2\right) , \tag{3.4}$$

where $\mathcal{H}f_q$ is the Hessian tensor of f at q.[2] A *Riemannian metric tensor* g on $\mathcal{P}$ is a symmetric, positive-definite (0,2)-tensor on $\mathcal{P}$. The value of the tensor g_q at point $q \in \mathcal{P}$ provides an inner product on the tangent space of $\mathcal{P}$ at q, and therefore specifies concepts such as distance, angle, and volume. We call the metric g *consistent* with the divergence $d_f(p,q)$ if

$$d_f(p,q) = g_q(p-q, p-q) + \mathcal{O}(\|p-q\|^2) ,$$

that is, g agrees with d up to quadratic terms in a neighborhood of q. From (3.9), it is clear that this consistency condition implies $g = \frac{1}{2}\mathcal{H}f$; that is, the Riemannian metric tensor is just a rescaling of the the Hessian tensor of f, as noted in [7]. Because $M \subset \mathcal{P}$, the restriction of g to M also endows M with the structure of a Riemannian submanifold of $\mathcal{P}$. Furthermore, the fact that M is parameterized allows us to define a pullback metric tensor $\alpha^* g$ on R via the formula

$$\alpha^* g_x(v_1, v_2) = g_{\alpha(x)}(D\alpha_x(v_1), D\alpha_x(v_2)) .$$

for any v_1, v_2 in the tangent space of M at x. We may express $\alpha^* g$ in local coordinates as the matrix function

$$\alpha^* g_x = \frac{1}{2} D\alpha_x^T \mathcal{H}f_{\alpha(x)} D\alpha_x . \tag{3.5}$$

The expression of $\alpha^* g$ in local coordinates allows us to quantify how the geometric properties of the information manifold M vary as we move through coordinate or "geographic" space R.

We note that the pullback tensor $\alpha^* g$ is a familiar object from parametric statis-

[2] The proof of (3.9) proceeds directly by computing a Taylor expansion and canceling out zeroth- and first-order terms.

tics. When $f(q) = \sum_i q_i \log q_i$, d_f is the Kullback-Leibler divergence, and $\alpha^* g$ may be calculated from (3.5) as

$$\begin{aligned}
\alpha^* g_x(e_i, e_j) &= \frac{1}{2}[D\alpha_x^T \mathcal{H} f_{\alpha(x)} D\alpha_x]_{ij} \\
&= \frac{1}{2} \sum_{y \in \mathcal{Y}} \left[\frac{\partial p(Y|x)}{\partial x_i} \frac{1}{p(Y|x)} \frac{\partial p(Y|x)}{\partial x_j} \right] \\
&= \frac{1}{2} \mathbb{E}_Y \left[\frac{\partial p(Y|x)}{\partial x_i} \frac{1}{p(Y|x)^2} \frac{\partial p(Y|x)}{\partial x_j} \middle| X = x \right] \\
&= \frac{1}{2} \mathbb{E}_Y \left[\frac{\partial \log p(Y|x)}{\partial x_i} \frac{\partial \log p(Y|x)}{\partial x_j} \middle| X = x \right] .
\end{aligned}$$

This last line defines the (expected) Fisher information matrix, a fundamental object in classical parametric statistical inference.

Not only is the pullback tensor $\alpha^* g$ familiar from parametric statistics; it is also related to a geometric localization of the mutual information.

Definition 7. *The* local information density *in X about Y at x_0 is*

$$j(x_0) \triangleq \lim_{r \to 0} \frac{I(X, Y | X \in B_r(x_0))}{r^2} , \tag{3.6}$$

where $B_r(x_0) = \{x \in R \mid \|x - x_0\|^2 \leq r^2\}$.

Intuitively, the local information density measures the degree of spatial variation in demographic variables per unit of spatial area. Theorem 2 ensures that the local information density exists everywhere, and gives a geometric interpretation in terms of the pullback metric.

Theorem 2. *The local information density $j(x)$ exists at all $x \in \mathrm{int}(R)$ and satisfies*

$$j(x) = \frac{1}{2} \frac{n}{n+2} \mathrm{tr}\,(\alpha^* g_x) \ . \tag{3.7}$$

While the proof is somewhat involved, the intuition behind Theorem 2 is simple to grasp. At the point x, the pullback metric $\alpha^* g_x$ is, in local coordinates, a matrix whose entries describe the degree to which the demographic distribution changes when one

moves along the principle axes in coordinate space. The average change when moving a unit distance away from x in a uniformly chosen direction is indeed $\operatorname{tr}(\alpha^* g_x)$. Both $j(x)$ and $\operatorname{tr}(\alpha^* g_x)$ therefore measure average change in the neighborhood of x, which similarity is made precise by Theorem 2.

The trace of $g_f(x)$ is the sum of its diagonal entries. Theorem 2 ensures that the local information density is well-defined, and provides a convenient way to compute it that does not depend on limiting operations.

Proof of Theorem 2

The proof proceeds essentially by direct calculation; in order to structure the calculation in a coherent way we divide it into a series of lemmas.

Recall that, by hypothesis, $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$, and therefore has a Radon-Nikodym derivative (probability density function) p_X, so that

$$\int f \, d\mathbb{P}_X = \int f p_X \, d\lambda$$

for any f. We also assumed that p_X is smooth, and will proceed to take derivatives and Taylor expansions accordingly.

Let $V_n(r)$ be the volume of the n-ball of radius r, and $S_{n-1}(r)$ the volume of the $n-1$-sphere of radius r. We will abbreviate $V_n = V_n(1)$ and $S_n = S_n(1)$. We have $V_n(r) = r^n V_n$ and $S_{n-1}(r) = r^{n-1} S_n$.

We use the standard symbol $o(f(r))$ to denote terms satisfying

$$\lim_{r \to 0} \frac{o(f(r))}{f(r)} = 0 \, .$$

Lemma 1. *We have*

$$\mathbb{P}(X \in B_r) = p_X(x_0) V_n(r) + o(r^{n+1}) \, .$$

Proof. We compute

$$\begin{aligned}\mathbb{P}(x \in B_r) &= \int_{B_r} d\mathbb{P}_X \\ &= \int_{B_r} p_X(x)\, d\lambda(x) \\ &= \int_{B_r} [p_X(x_0) + Dp_X(x_0)(x - x_0) + o(\|x - x_0\|)\,]\, d\lambda(x)\,.\end{aligned}$$

The middle term in the integral vanishes via spherical symmetry, and we obtain

$$\begin{aligned}\mathbb{P}(x \in B_r) &= \int_{B_r} [p_X(x_0) + o(\|x - x_0\|)\,]\, d\lambda(x) \\ &= p_X(x_0)V_n(r) + \int_{B_r} o(\|x - x_0\|)\, d\lambda(x) \\ &= p_X(x_0)V_n(r) + \int_{B_r} o(r)\, d\lambda(x) \\ &= p_X(x_0)V_n(r) + o(r^{n+1})\,,\end{aligned}$$

as was to be shown. □

For notational convenience, let $\alpha(x) = p(\cdot|x)$ and $a = p(\cdot|X \in B_r)$.

Lemma 2. *We have*

$$a = \alpha(x_0) + o(r)\,.$$

Proof. We compute

$$\begin{aligned}a &= \mathbb{E}[\alpha(X)|X \in B_r] \\ &= \int \alpha(x)\, d\mathbb{P}_{X|X\in B_r}(x) \\ &= \int \alpha(x)p(x|X \in B_r)\, d\lambda(x) \\ &= \frac{1}{\mathbb{P}(X \in B_r)} \int_{B_r} \alpha(x)p(x)\, d\lambda(x)\,.\end{aligned}$$

Since $\alpha(x)$ and $p(x)$ are both smooth, their product is as well, and we may Taylor expand about $x = x_0$ to obtain

$$a = \frac{1}{\mathbb{P}(X \in B_r)} \int_{B_r} [\alpha(x_0)p(x_0) + T(x - x_0) + o(\|x - x_0\|)]\, d\lambda(x)\,,$$

where T stands for a linear map that we need not calculate, since this term vanishes in the integral through spherical symmetry. Thus,

$$\begin{aligned} a &= \frac{1}{\mathbb{P}(X \in B_r)} \int_{B_r} [\alpha(x_0)p(x_0) + o(\|x - x_0\|)]\, d\lambda(x) \\ &= \frac{\alpha(x_0)p(x_0)V_n(r) + o(r^{n+1})}{\mathbb{P}(X \in B_r)} \\ &= \frac{\alpha(x_0)p(x_0)V_n(r) + o(r^{n+1})}{p(x_0)V_n(r) + o(r^{n+1})} \qquad \text{(Lemma 1)} \\ &= \alpha(x_0) + o(r)\ , \end{aligned}$$

as was to be shown.

□

Lemma 3. *For any $x \in B_r$,*

$$d_f(p(\cdot|x), p(\cdot|X \in B_r)) = d_f(p(\cdot|x), p(\cdot|x_0)) + o(r^2)\ . \tag{3.8}$$

Proof. The proof proceeds by exploiting the local quadratic structure of the Bregman divergence d_f:

$$d_f(q + \delta, q) = \frac{1}{2}\mathcal{H}f_q(\delta, \delta) + o\left(\|\delta\|^2\right)\ , \tag{3.9}$$

Using the same notations $\alpha(x) = p(\cdot|x)$ and $a = p(\cdot|X \in B_r)$, we compute

From 3.9, we obtain

$$\begin{aligned} d_f(p(\cdot|x), p(\cdot|X \in B_r)) &= d_f(\alpha(x), a) \\ &= \mathcal{H}f_a(\alpha(x) - a, \alpha(x) - a) + o(\|\alpha(x) - a\|^2)\ . \end{aligned}$$

Since $p(y|x)$ is smooth as a function of x, so is $\alpha(x)$, and the final term is therefore $o(\|r\|^2)$. Rearranging terms, we can write

$$\begin{aligned} d_f(p(\cdot|x), p(\cdot|X \in B_r)) = \mathcal{H}f_{\alpha(x)}(\alpha(x) - a, \alpha(x) - a) + \\ (\mathcal{H}f_a - \mathcal{H}f_{\alpha(x)})(\alpha(x) - a, \alpha(x) - a) + o(\|\alpha(x) - a\|^2\ . \end{aligned}$$

Since f is smooth, the components of the tensor $\mathcal{H}f_a - \mathcal{H}f_{\alpha(x)}$ are $o(r)$. Furthermore, by Lemma 2, $\alpha(x) - a = \alpha(x) - \alpha(x_0) + o(r) = O(r) + o(r)$. The entire second term is therefore $o(r^2)$. Turning to the first term, we have

$$\begin{aligned}\mathcal{H}f_{\alpha(x)}(\alpha(x)-a,\alpha(x)-a) &= \mathcal{H}f_{\alpha(x)}(\alpha(x)-\alpha(x_0)+o(r),\alpha(x)-\alpha(x_0)+o(r)) \\ &= \mathcal{H}f_{\alpha(x)}(\alpha(x)-\alpha(x_0),\alpha(x)-\alpha(x_0)) + o(r^2) \\ &= d_f(\alpha(x),\alpha(x_0)) + o(r^2)\ ,\end{aligned}$$

where in the second line we have used the fact that $\alpha(x) - \alpha(x_0) = O(r)$. This completes the proof. □

Lemma 4. *Let T be a real, symmetric bilinear form, and Δ_2 the diagonal operator. Then,*

$$\int_{B_r(0)} T \circ \Delta_2 \, d\lambda = \frac{r^{n+2} S_{n-1}}{n(n+2)} \mathrm{tr}\,(T)\ .$$

Proof. By the spectral theorem, there exists an orthonormal basis in which the matrix A of T is diagonal, and its entries are the eigenvalues $\{\lambda_i\}$ of T. Since $B_r(0)$ is radially symmetric about the origin, we may integrate in this basis instead, obtaining

$$\int_{B_r(0)} T \circ \Delta_2 \, d\lambda = \int_{B_r(0)} v^T A v \, d\lambda(v)\ .$$

Since A is diagonal,

$$\begin{aligned}\int_{B_r(0)} v^T A v \, d\lambda(v) &= \int_{B_r(0)} \sum_i \lambda_i v_i^2 \, d\lambda(v) \\ &= \sum_i \lambda_i \int_{B_r} v_i^2 \, d\lambda\ .\end{aligned}$$

By spherical symmetry, the integrals inside the sum are all equal to $\frac{1}{n}\int_{B_r(0)} \|v\|^2 \, d\lambda(v)$,

and we obtain

$$\begin{aligned}
\sum_i \lambda_i \int_{B_r} v_i^2 \, d\lambda &= \frac{(\sum_i \lambda_i)}{n} \int_{B_r(0)} \|v\|^2 \, d\lambda(v) \\
&= \frac{\operatorname{tr}(T)}{n} \int_{B_r(0)} \|v\|^2 \, d\lambda(v) \\
&= \frac{\operatorname{tr}(T)}{n} \int_{\rho \in [0,r]} \rho^2 S_{n-1}(\rho) \, d\rho \qquad \text{(polar coordinate transform)} \\
&= \frac{\operatorname{tr}(T)\, S_{n-1}}{n} \int_{\rho \in [0,r]} \rho^{n+1} \, d\rho \\
&= \frac{S_{n-1} r^{n+2}}{n(n+2)} \operatorname{tr}(T) \ ,
\end{aligned}$$

as was to be shown. □

We are now prepared to prove the theorem. We have

$$\begin{aligned}
I(X, Y | X \in B_r) &= \int_{B_r} d_f(p(\cdot|x), p(\cdot|X \in B_r)) \, d\mathbb{P}_{X|X \in B_r}(x) \\
&= \int_{B_r} d_f(\alpha(x), a) \, d\mathbb{P}_{X|X \in B_r}(x) \\
&= \int_{B_r} d_f(\alpha(x), a) p(x|X \in B_r) \, d\lambda(x) \\
&= \frac{1}{\mathbb{P}(X \in B_r)} \int_{B_r} d_f(\alpha(x), a) p(x) \, d\lambda(x) \\
&= \frac{1}{p(x_0) V_n(r + o(r^{n+1}))} \int_{B_r} d_f(\alpha(x), a) p(x) \, d\lambda(x) \ . \qquad \text{(Lemma 1)}
\end{aligned}$$

Focusing now on the integral, we have

$$\begin{aligned}
\int_{B_r} d_f(\alpha(x), a) p(x) \, d\lambda(x) &= \int_{B_r} \left[d_f(\alpha(x), \alpha(x_0)) + o(r^2) \right] p(x) \, d\lambda(x) \\
&= \frac{1}{2} \int_{B_r} \left[\mathcal{H} f_{\alpha(x_0)}(\alpha(x) - \alpha(x_0), \alpha(x) - \alpha(x_0)) + o(r^2) \right] p(x) \, d\lambda(x) \ .
\end{aligned}$$

Since $\alpha(x) - \alpha(x_0) = D\alpha_{x_0}(x - x_0) + o(r)$, we can rewrite the integrand, obtaining

$$\int_{B_r} d_f(\alpha(x), a) p(x) \, d\lambda(x) = \int_{B_r} \left[(\mathcal{H} f_{\alpha(x_0)} \circ \Delta_2 \circ D\alpha_{x_0})(x - x_0) + o(r^2) \right] p(x) \, d\lambda(x)$$

Since $p(x)$ is smooth, $p(x) - p(x_0) = o(r)$, and we can further rewrite the integral as

$$\begin{aligned}\int_{B_r} d_f(\alpha(x), a)p(x)\, d\lambda(x) &= \int_{B_r} \left[(\mathcal{H}f_{\alpha(x_0)} \circ \Delta_2 \circ D\alpha_{x_0})(x - x_0) + o(r^2)\right] p(x_0)\, d\lambda(x) + o(r^{n+2}) \\ &= p(x_0) \int_{B_r} \left[(\mathcal{H}f_{\alpha(x_0)} \circ \Delta_2 \circ D\alpha_{x_0})(x - x_0)\right]\, d\lambda(x) + o(r^{n+2}) \\ &= p(x_0) \int_{B_r(0)} (\mathcal{H}f_{\alpha(x_0)} \circ \Delta_2 \circ D\alpha_{x_0})\, d\lambda + o(r^{n+2})\end{aligned}$$

By Lemma 4,

$$\begin{aligned}\int_{B_r(0)} (\mathcal{H}f_{\alpha(x_0)} \circ \Delta_2 \circ D\alpha_{x_0})\, d\lambda &= \frac{S_{n-1} r^{n+2}}{n(n+2)} \mathrm{tr}\left(\mathcal{H}f_{\alpha(x_0)} \circ \Delta_2 \circ D\alpha_{x_0}\right) \\ &= \frac{S_{n-1} r^{n+2}}{n(n+2)} \mathrm{tr}\left(g_{x_0}\right) .\end{aligned}$$

Returning now to the main calculation, we have

$$I(X, Y | X \in B_r) = \frac{1}{p(x_0)V_n(r) + o(r^{n+1})} \left[p(x_0) \frac{1}{2} \frac{S_{n-1} r^{n+2}}{n(n+2)} \mathrm{tr}\left(g_{x_0}\right) + o(r^{n+2}) \right] .$$

We can cancel terms, using the fact that $S_{n-1}/V_n = n$, obtaining

$$I(X, Y | X \in B_r) = \frac{1}{2} \frac{r^2}{n+2} \mathrm{tr}\left(g_{x_0}\right) + o(r^2) .$$

Dividing through by r^2 and taking the limit as $r \to 0$ proves the theorem.

Geodesic Distances

Finally, we note that the pullback metric $\alpha^* g$ provides a standard way to measure distances on $\mathcal{M}$.[3]

Definition 8. *The geodesic distance $\delta(x_1, x_2)$ between points $x_1, x_2 \in R$ is*

$$\delta(x_1, x_2) = \min_{\gamma \in C(x_1, x_2)} \int_0^1 \sqrt{\alpha^* g_{\gamma(t)}(\gamma'(t), \gamma'(t))}\, dt ,$$

[3]The geodesic distance is normally defined between points on the manifold M, not points in the coordinate space R. Because R parameterizes M via a single chart α, our definition is equivalent to the standard one.

where $C(x_1, x_2)$ is the set of all unit-speed curves in R joining x_1 to x_2.

Recalling that $\alpha^* g$ encodes the relationship between spatial change and demographic change, we can interpret (8) as the minimal amount of demographic change required to travel on a continuous path from point x_1 to point x_2. Regions with similar demographics may be far apart according to the geodesic distance if they are separated by a region with very different demographics.

We pause to summarize the mathematical development thus far. We have formulated a mathematical space Γ of compositional triples, consisting in a geographic space R, a population measure μ, and a map α relating space and demographics. For each such triple γ we obtain an information manifold $\mathcal{P}$ that encodes both spatial and demographic information. By choosing a Bregman divergence d_f, we obtain a metric tensor g on $\mathcal{P}$ and a pullback metric $\alpha^* g$ on R which record how geographic distances in R are related to information distances on $\mathcal{P}$. The pullback metric $\alpha^* g$ is closely related to the local information density via (3.7), and allows the explicit computation of information distances. In the next section, we will consider the problem of *dimension reduction* in the Bregman formalism, through the framework of regionalization.

3.2.5 Learning the Structure of Segregation

Operationally, we view the problem of learning the structure of segregation as the task of finding interpretable units of spatial aggregation whose boundaries correspond to demographic transitions. This problem is a form of *regionalization* – spatially-constrained clustering. While recent papers have developed an array of methods for regionalization [38, 66, 82, 219], none are designed for multiscale segregation studies. On the other hand, a recent ethnoracial neighborhood identification from within segregation studies [193] does not produce contiguous spatial units. These limitations motivate our development of new methods.

Let $c : \mathcal{X} \to \{1, \ldots, k\}$ be a function that assigns to each location x a region label $c(x)$. We regard $C = c(X)$ as a random variable, and aim to choose c such that the aggregation it induces captured segregation at large spatial scales. The Chain Rule

of Bregman information [65] offers a decomposition of the form

$$I_f(X,Y) = I_f(C,Y) + I_f(X,Y|C)\ . \tag{3.10}$$

This result is often derived for the bespoke measures surveyed in Section 4.1, and often referred to as "decomposability." The generality of (3.10) explains why so many different measures share the same property – they are all instantiations of the Chain Rule.

The term $I_f(C,Y)$ gives the segregation captured at the aggregate spatial scale, and $I_f(X,Y|C)$ the residual segregation at lower scales. A good labeling function will tend to make the first term large. This motivates the following heuristic problem:

$$\begin{aligned} c^* = &\text{argmax}_c \quad I_f(c(X),Y) \\ &\text{subject to} \quad \text{spatial constraints.} \end{aligned} \tag{3.11}$$

In this paper we will specify the spatial constraints in terms of contiguity and soft regularity requirements, but other spatial constraint formulations may be desirable in other contexts. For most reasonable specifications of the spatial constraints, (3.11) is not efficiently solvable, and so we now develop two approximate methods.

Our first method is a form of greedy information maximization. We begin with a complete set of unclustered tracts. At each stage, we merge a pair of tracts or clusters together to form a single larger cluster, whose demographic counts are just the sum of the counts of the component pair. The merged cluster inherits the adjacencies of each of its components. We chose which tracts or clusters to merge together by attempting to greedily maximize the mutual information $I(C,Y)$ between cluster labels C and demographics Y. This may be achieved by merging the clusters i^* and j^* that satisfy

$$(i^*,j^*) = \underset{(i,j)\in\mathcal{E}(R)}{\text{argmin}}\ d_{js}(x_i,x_j), \tag{3.12}$$

where d_{js} is the asymmetric Jensen-Shannon divergence given by[4]

$$d_{js}(x_i, x_j) = \mu(x_i) d_f(\alpha(x_i), q_{ij}) + \mu(x_j) d_f(\alpha(x_j), q_{ij}) \ ,$$

where

$$q_{ij} = \frac{\mu(x_i)\alpha(x_i) + \mu(x_j)\alpha(j)}{\mu(\{x_i, x_j\})} \ .$$

It can be shown via direct calculation that $d_{js}(i, j)$ is the reduction in mutual information caused by merging i with j. The constraint $(i, j) \in \mathcal{E}(R)$ in (3.12) ensures that only adjacent clusters are merged together. Algorithm 1 formalizes the algorithm.

Algorithm 1: Greedy Agglomerative Information Maximization

Input: Set R of locations, demographic map p, desired number of clusters k

1 **while** $|R| > k$ **do**

2 $\quad i^*, j^* \leftarrow \operatorname{argmin}_{(i,j)\in\mathcal{E}(R)} d_{js}(x_i, x_j)$ $W \leftarrow W \setminus R$

3 $\quad$ aggregate(i, j) ▷ `Reduces` $|R|$ `by 1.`

Output: R

This algorithm, a form of agglomerative clustering, has the virtues of computational performance and direct optimization of the between-clusters Bregman information. Additionally, as we will see, the explicitly hierarchical structure of agglomerative clustering enables useful forms of data analysis and visualization, as we will see in Section 3.3.3.

In some cases, pure greedy clustering can lead to groups of highly irregular shape. Preprocessing the data in order to coarse-grain it can serve as a regularizer which avoids this behavior. While multiple methods are possible, we turn to spectral clustering [188, 130]. Spectral methods are well suited for regionalization [219], as they approximately solve an *normalized cut* problem that is explicitly formulated in terms of spatial boundaries. Spectral clustering proceeds by constructing a matrix of similarities between data points, whose spectrum is then used to carry out the clustering. We define a sparse distance matrix in which each node is connected to its neighbors

[4]We are using a slight abuse of notation by using μ to refer to the *discrete* population measure on the observed data, which allows each x_i to have nonzero measure.

in the adjacency network of Figure 3-4b. The edge-weights for this computation are just the distances between adjacent nodes, computed as

$$\bar{\delta}(x_i, x_j) = \sqrt{\frac{1}{2}(\alpha(x_i) - \alpha(x_j))^T H f_{p_{ij}}(\alpha(x_i) - \alpha(x_j))}\,, \tag{3.13}$$

where $p_{ij} = (\alpha(x_i) + \alpha(x_j))/2$.

We now state our algorithm for spectral clustering. Spectral clustering proceeds by constructing a graph Laplacian matrix L in whose eigenspace clusters are approximately separable, and may therefore be determined by the k-means algorithm.[5] The algorithm has three inputs: the region R endowed with adjacency network structure; a tunable parameter σ that controls the bandwidth of a Gaussian radial basis function used to compute the graph Laplacian; and a desired number of clusters k.

Algorithm 2: Attribute-Based Spectral Clustering

Input: Set R of locations, smoothing parameter σ, desired number of clusters k

for $x_i, x_j \in R$ **do**
  $A_{ij} \leftarrow \exp\left[-\frac{\delta(x_i,x_j)^2}{2\sigma}\right]$ ▷ `Gaussian RBF affinity matrix;` $\delta(x_i, x_j)$ `as in` (8)
$D \leftarrow \mathrm{diag}(e^T A)$ ▷ `Diagonal matrix of weighted node degrees`
$L \leftarrow D^{-1}(D - A)$ ▷ `Random walk normalized graph Laplacian`
$V \leftarrow [v_1, \ldots, v_k]$ ▷ v_ℓ `the` ℓ`th principal eigenvector of` L

Output: $\mathrm{kmeans}(V, k)$

3.3 Results

In this section, we develop computational tools for studying widely-available data sets using information-geometric concepts. Our primary aim is to develop practice-ready computational methods for our two main tasks of quantifying spatial scale and learn-

[5] There are multiple definitions of the Laplacian matrix. The one we use below is the "random-walk normalized Laplacian," the same one used in the earliest use of spectral clustering by [188]. Readers interested in the properties of classical spectral clustering are encouraged to explore the review [130].

ing spatial structure. First, we describe the data sets we will study, and show how they may be approximated as information manifolds. We then show how to compute the metric tensor from data. Finally, we formulate two algorithms – one spectral and one hierarchical – for exploratory clustering of cities based on demographic attributes. Throughout our development, we illustrate our methods in the context of racial residential segregation in major U.S. cities. In all analyses conducted in this section, we choose our generating function f to be the negative Shannon entropy.

3.3.1 Data Access and Structure

Our data is extracted from table B03002 of the 5-year American Community Survey, end year 2013, entitled "Hispanic or Latino Origin By Race."[6] From this table, we aggregated racial and ethnic categories into five supercategories, "Asian," "Black," "Hispanic," "Other," and "White."[7] The retrieved data consist of a set of polygons ("tracts"),[8] as well as an estimated count of the number of residents of different demographic groups within each tract (Figure 3-3). The delimiting of cities based on non-arbitrary population densities or natural boundaries is an area of active discussion in urban planning [179, 180]. Our simple approach in this study was to analyze the region composed of all counties in which some or all of the city's municipal boundaries lie.

Information geometry requires differentiable structure, but the set of tract polygons is discrete. Figure 3-4 illustrates how we approximate the information manifold M using such data. From raw tracts (Figure 3-4a) we construct the tract adjacency matrix (Figure 3-4b). The attribute map α represents nodes as points in the probability simplex $\mathcal{P}$, as shown in Figure 3-4c, while the edges between them give topological information necessary for the computation of the metric tensor and other geometric quantities.

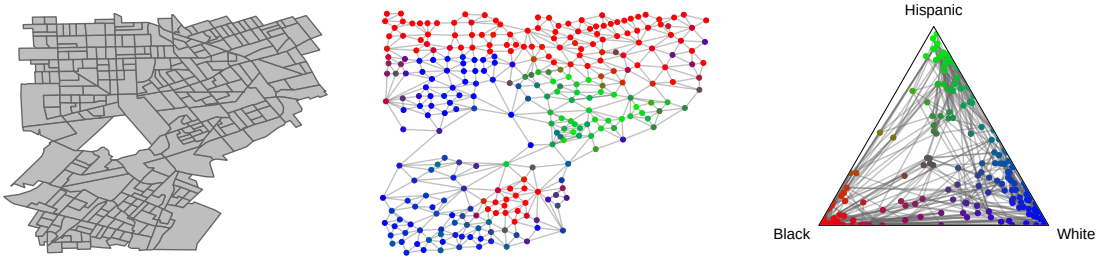
[6] These data were retrieved directly from the Census Bureau API using the `tigris` and `acs` packages for the `R` programming language [209, 89].

[7] For convenience, we will refer to these as "racial" groups, not withstanding continued discussion over whether Hispanic origin is correctly viewed as a racial characteristic.

[8] We will colloquially refer to the polygons as tracts, despite the fact that the data we used is the higher-resolution Census blockgroup.

Tract ID	Asian	Black	Hispanic	Other	White
261635016001	6	835	52	42	213
261635017001	11	819	17	38	211
261635061002	26	320	34	11	39

Figure 3-3: Excerpt of demographic data in Detroit, after aggregation into five racial groups.
The tract ID specifies the polygon to which the demographic information corresponds.



(a) ACS Blockgroups (b) R (c) $M \subset \mathcal{P}$

Figure 3-4: Data structure and representation.
3-4a: Tracts corresponding to a small subsection of Detroit. 3-4b: Adjacency network of tracts, representing the topological structure of R. The nodes are colored reflecting their demographic composition, with blue representing white residents, red representing black residents, and green representing Hispanic residents. Asian and Other residents are not shown. 3-4c: Image of R under the attribute map α.

3.3.2 Local Information and the Scale of Segregation

The pullback metric $\alpha^* g_x$ is required for both the computation of the local information density per (3.7) and the geodesic distance per (8). Equation (3.5) indicates that we may compute the pullback metric by estimating the derivative of the attribute map $D\alpha$ at x and computing the Hessian tensor $\mathcal{H}f$ at x. For most common choices of f the Hessian may be computed analytically, so we focus on the computation of the derivative $D\alpha$. Since the direct computation of difference quotients on observed data typically leads to numerical instability, we instead use a more robust method based on weighted linear regression. We regress the attribute differences $\alpha(x_i) - \alpha(x)$ on the geographic displacements $x_i - x$; the regression coefficients then approximate the components of the derivative $D\alpha$. Let $\mathcal{E}(x)$ denote the ego network of node x in the contiguity network of analytical units as shown in Figure 3-4b; larger neighborhoods can be specified in order to smooth the results. Our approximation formula is then

$$D\alpha_x \approx (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}_x \,, \tag{3.14}$$

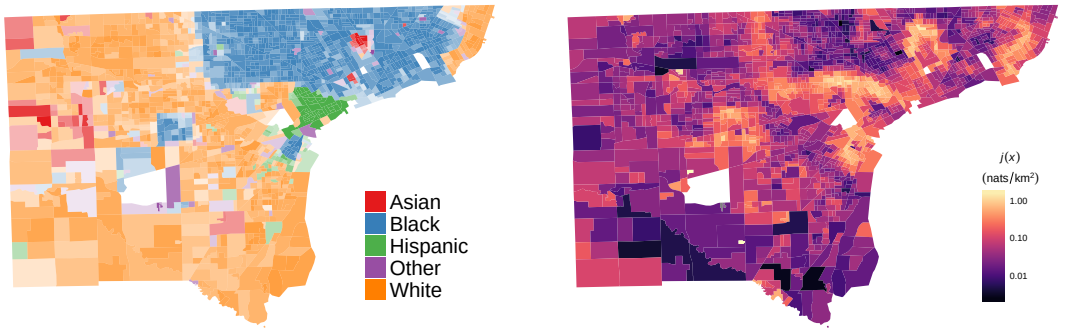
where:

- $\mathbf{X}$ is the matrix whose ith row is the difference $x_i - x$ for each $x_i \in \mathcal{E}(x)$.
- $\mathbf{W}$ is a diagonal weighting matrix that prioritizes tracts closer to the origin x. We used a Gaussian radial basis weight, yielding

$$\mathbf{W}_{ij} = \begin{cases} \exp\left[-\frac{\|x_i - x\|^2}{2\sigma}\right] & i = j \\ 0 & \text{otherwise} \,, \end{cases}$$

  where σ is a tunable characteristic length scale set to 10km in our computations, corresponding to very weak weighting.
- $\mathbf{Y}_x$ is the matrix whose ith row is the vector $\alpha(x_i) - \alpha(x)$.

With $D\alpha_x$ and $\mathcal{H}f_x$ in hand, the pullback metric $\alpha^* g_x$ may then be computed via (3.5).



(a) (b)

Figure 3-5: Illustration of the local information density.
3-5a: Ethnoracial prominence map in Detroit. 3-5b: Estimates of the local information density, computed according to (3.5) and (3.14). A Gaussian smoothing kernel was applied to the demographic data prior to the computation of $j(x)$ for visualization purposes.

As a first application, we illustrate the computation the local information density on our sample cities. Figure 3-5 gives an example computation of the local information density in Wayne County, Michigan, which includes the city of Detroit. In Figure 3-5a, we highlight geographic patterns of racial difference. The color of each tract reflects the group $i^* = \operatorname{argmax}_i r(i) \triangleq \operatorname{argmax} \alpha(x)_i \log \frac{\alpha(x)_i}{\bar{q}_i}$, where $\bar{q}$ is the global demographic distribution, and therefore reflects groups that are most locally overrepresented relative to the city's population. The saturation is proportional to $r(i^*)$. In Figure 3-5b, we compute and visualize the local information density $j(x)$ in each of the tracts. We observe that $j(x)$ is highest in areas of demographic transition, such as the border between the white western suburbs and the black urban core toward the east. Other transitional areas with high local information density include the primarily white and Asian town of Hamtramck toward the northeast, and the predominantly black town of Inkster lying within the white suburbs to the west. These results illustrate the role of the local information density in highlighting boundaries between regions of highly varying demographic compositions.

In addition to visualizing regions of spatial transition, the local information density may also be used to characterize the average scale of spatial segregation in cities for comparative purposes.

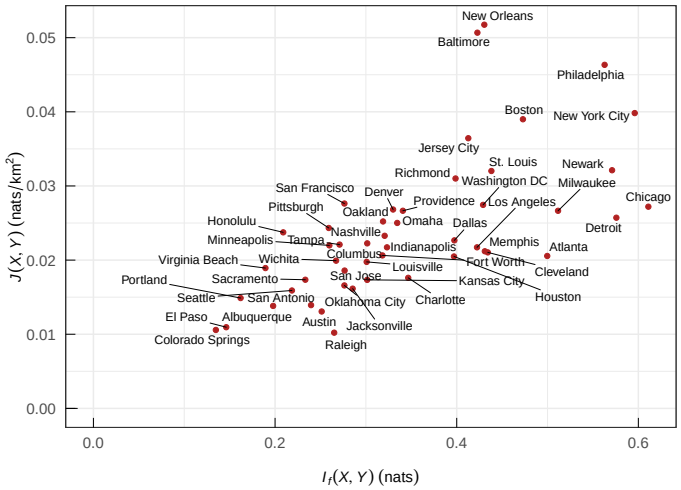


Figure 3-6: Segregation profiles for 56 major U.S. cities.
The mutual information $I(X,Y)$ quantifies the overall degree of spatial separation, while the mean local information density $J(X,Y)$ reflects the average spatial scale of separation.

Definition 9. *The* mean local information density *in* X *about* Y *is*

$$J(X,Y) = \int_R j \, d\mu \, .$$

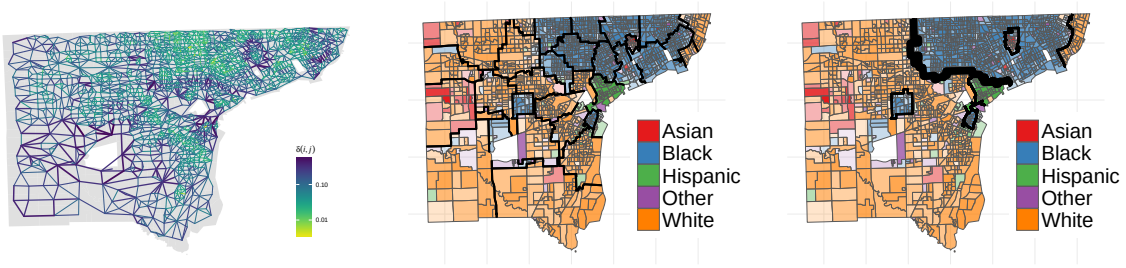
The mean local information density $J(X,Y)$ may be directly computed from the values of $j(x)$, with normalized population proportions supplying the measure μ. Cities with mean information density $J(X,Y)$ are those with many spatial transitions, implying smaller, densely-packed demographic regions. In contrast, cities where $J(X,Y)$ is low have few spatial transitions, indicating that racial difference exists on the scale of large, demographically homogeneous mega-regions. The mean local information density thus provides a principled solution to the checkerboard problem that does not require the analyst to set a potentially arbitrary choice of analytical scale. Cities whose layout resembles the binary city of Figure 3-1a will have low information densities due to the presence of a small number of boundaries separating large, monolithic regions. Cities whose layout more resembles the checkerboard of Figure 3-1b, on the other hand, will have much larger densities reflecting the presence of more boundaries between regions.

Figure 3-6 plots segregation profiles of 56 study cities. The (aspatial) mutual information $I(X,Y)$ captures the overall degree of spatial separation in the city, while the mean local information density $J(X,Y)$ measures the spatial scale on which the separation exists. Cities where $I(X,Y)$ is low are relatively unsegregated, either because they are nearly monoracial or because racial groups in those cities tend to coexist in the same areas. As $I(X,Y)$ increases toward the right of the figure, a spectrum of segregation structures emerge. Cities such as Detroit, Atlanta, and Chicago are sharply segregated into megaregions; these concentrate toward the bottom-right of the plot. Cities on the top-right such as New York and Philadelphia are also sharply segregated, but at much smaller characteristic spatial scales. These results are directionally aligned with the similarly-motivated segregation ratio of [172]; however, the segregation ratio does not provide any local information concerning the dependence of spatial scale on location. The local information density is therefore to be preferred for detailed study of the structure of individual cities.

3.3.3 Decomposition and the Structure of Segregation

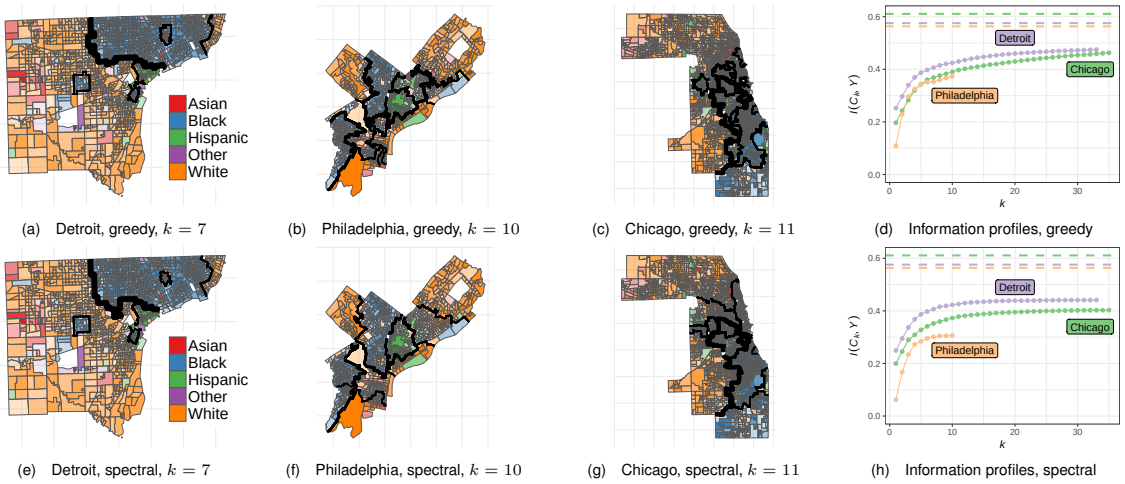
Figure 3-7b illustrates the complete pipeline, from weighted distance matrix to a spectral intermediate partition to a final hierarchical partition. Figure 3-7a visualizes the distance network used for spectral clustering, with distances computed according to (3.13). Darkly shaded edges in this network reflect adjacent nodes with large geodesic distance between them as calculated by (3.13), indicating that the two nodes are quite demographically different. Figure 3-7b shows an illustrative spectral partition into 30 regions, overlaid on the demographic map. Finally, Figure 3-7c shows the result of greedily agglomerating the 30 spectral regions into six composite regions.

Figure 3-8 gives a comparative analysis of both pure greedy and spectrally-preprocessed regionalizations in Detroit, Chicago, and Philadelphia. First, the plot of information captured against number of regions shown in Figure 3-8(d) provides a profile curve of segregation against the number of clusters. Unlike previous profile methods, these curves do not assume the scale to be global at any of its points, and support mathematically precise decomposition claims about how much segregation is captured at a



(a) Distance matrix (b) Spectral intermediate partition. (c) Hierarchical partition.

Figure 3-7: Illustration of the regionalization pipeline in Wayne County, Michigan. 3-7a: The adjacency network. Darker edges depict "obstructions" in the network in which the components of the distance matrix are large. 3-7b: Spectral clustering of this network into 30 regions, overlaid on the demographic map. 3-7c: Hierarchical clustering of the 30 spectral regions.



(a) Detroit, greedy, $k = 7$ (b) Philadelphia, greedy, $k = 10$ (c) Chicago, greedy, $k = 11$ (d) Information profiles, greedy

(e) Detroit, spectral, $k = 7$ (f) Philadelphia, spectral, $k = 10$ (g) Chicago, spectral, $k = 11$ (h) Information profiles, spectral

Figure 3-8: Illustrative partitionings of Detroit, Chicago, and Philadelphia under pure greedy (top) and greedy with spectral preprocessing regionalization methods (bottom).

The value of k gives the number of partitions, and the weight of each boundary reflects its contribution to overall segregation. In Figures 3-8(e)-(g), $\sigma = 30$ was used for all spectral preprocessing. Figures 3-8(d) and 3-8(h) give profile curves under each method. The dashed lines give the total segregation $I_f(X, Y)$.

given scale. Furthermore, we also obtain a particularly powerful form of decomposition analysis. The agglomerative method has an attractive structural feature: to each boundary between regions, agglomerative clustering assigns an information value reflecting its contribution to overall segregation. The sum of all such information values is overall segregation $I_f(X, Y)$. Furthermore, we also obtain a particularly powerful form of decomposition analysis, in which we can decompose overall segregation into terms corresponding to concrete spatial features. In Figure 3-8(a), for example, agglomerative clustering highlights the dividing line between the predominantly black urban core of Detroit and the predominantly white suburbs; this single boundary accounts for a full 44% of segregation in Detroit. The hierarchical nature of the algorithm additionally specifies smaller subdivisions nested within this dominating boundary. The seven regions shown account for 71% of total segregation.

Figures 3-8(b)-(c) also highlight some characteristic limitations of greedy agglomerative methods. First, because greedy partitioning requires nothing more than a connectivity constraint between tracts, the regions it finds may be highly irregular in shape, as seen in Chicago and especially Philadelphia. Whether this is acceptable depends on the analytical context. Second, greedy partitioning is extremely sensitive to data perturbations, which can leads to unpredictable results: in Chicago, for example, the small region to the northeast does not appear sharply differentiated from the surrounding, predominantly white suburbs. However, the region to the west combines black and Hispanic neighborhoods, suggesting that the algorithm has missed a natural spatial boundary. These limitations are largely addressed by the incorporation of spectral preprocessing, shown in Figures 3-8(e)-(g). In Detroit, the spectral preprocessing finds substantively identical regions to those found through pure greed. In Chicago and Philadelphia, spectral preprocessing results in regions with more regular shapes, and distinguishes, for example, the western Hispanic and black regions missed by pure greedy partitioning in Chicago. On the other hand, information values for these partitionings are slightly lower than the corresponding greedy partitions; the parameters discussed above allow the analyst to exercise control over this tradeoff.

These methods extend easily to the study of boundaries in space as well as time.

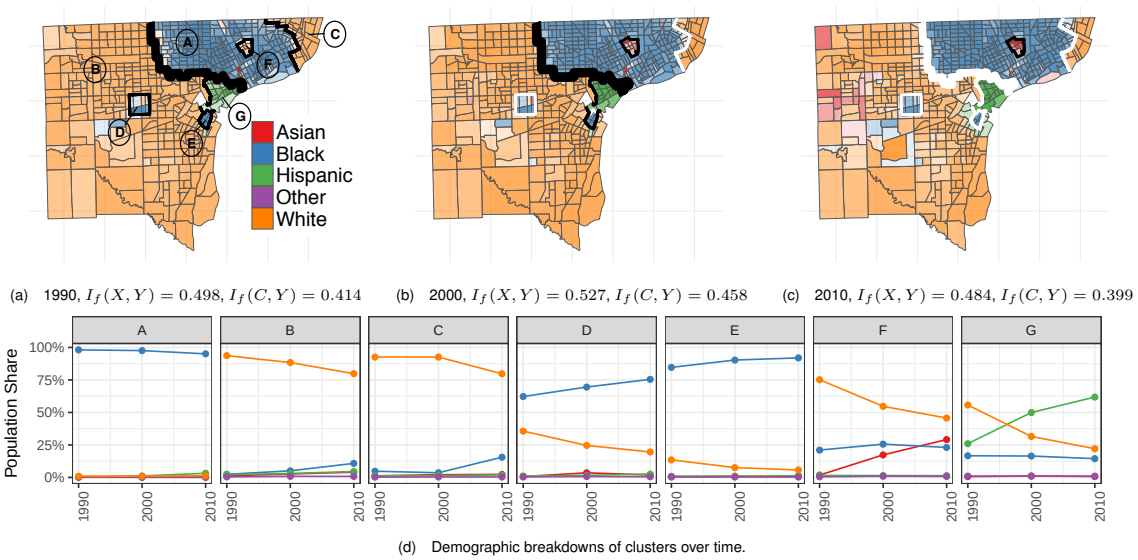


Figure 3-9: Evolving spatial boundaries in Detroit.
The size of each spatial boundary corresponds to its information value as identified via greedy partitioning with spectral preprocessing. In Figure 3-9b-c, white boundaries have decreased in segregation magnitude and black ones have increased relative to the previous time-step. The demographic breakdowns in Figure 3-9d correspond to the region labels of Figure 3-9a. In these figures only, data was used on a tract level as supplied by [128].

This may be achieved by connecting time-slices into a single, temporal network on which we perform regionalization. As an example, Figure 3-9 shows the dynamics of segregation in Detroit over the time-period 1990-2010, data for which we obtained from [128] to obtain consistent spatial units across time. Figures 3-9(a)-(c) show evolving demographics, as well as $k = 7$ labeled spatiotemporal regions obtained via spectral partitioning with hierarchical post-processing. The present methods allow us to directly read off the changing shape and contribution to segregation of spatial boundaries. From 1990 to 2000, overall segregation increased in Detroit. This largely reflects increasing isolation of suburban whites (region B) from blacks and Hispanics, but within this overall trend are easily-missed nuances. These nuances are shown in the figure by black boundaries along which segregation increased, and white ones along which it decreased, relative to the previous time-step. For example, separation between the Grosse Point communities (cluster C) and the predominantly black cluster A decreased. Indeed, the analysis suggests that segregation in and around Grosse

Point changed rapidly in this time period; a cluster of tracts that were demographically more similar to Grosse Point in 1990 more closely resemble central Detroit in 2000, and the boundary between the two clusters shifts accordingly. Overall segregation decreased in 2010, largely due to amelioration of the spatial divisions that had intensified in the previous decade. However, the growing Asian population of Hamtramck (cluster F) increasingly distinguished it from its surroundings. Regionalization allows us to quantify not only *that* segregation has evolved in Detroit, but also *how* it has evolved.

In practice, we are not given whether a given data set is regionalizable. The local information density discussed below or other exploratory methods can be used to determine whether a given data set is indeed "regionalizable" into spatially distinct demographic regions. The analyst must then choose a desired number of final clusters, and, in the case of spectral preprocessing, the hyperparameter σ and number of intermediate clusters. A simple approach to selection problem is to fix a desired number of final clusters, and then conduct grid search over σ and the number of intermediate clusters, using the mutual information as a loss function. More detailed approaches involving the inspection of the spectrum of the graph Laplacian are also possible.

3.4 Discussion

We have developed a suite of methods for studying the structure of spatial segregation using modern information theory and machine learning. These methods advance profile curve approaches by generating curves that do not assume global scale and that support decomposition claims. They advance decomposition methods by constructing non-arbitrary boundaries based on spatial demographic trends, which may be studied directly or used as input to further analysis. The local information density may be averaged and compared between cities, but, unlike existing scale measurements, it may also be used for detailed analysis to study how characteristic spatial scales of segregation vary across the region of analysis. In sum, these advances enable a range

of analyses not previously possible in a systematic, scalable fashion.

While our example has been ethnoracial segregation, these methods generalize in two important ways. First, they are not limited to demographic study; arbitrary compositional data with spatial correlations may be used, such as frequently arise in ecology, geology, and geography. Second, they are not limited to categorical variables; the formalism of Bregman information allows the use of these methods for ordinal and partially ordinal variables as well. Many other generalizations and modifications are possible. It may, for example, be of interest to conduct soft-regionalization, in which each spatial unit is assigned a fractional membership score for each region. Such analysis may be especially appropriate when distinct spatial regions are separated by soft gradients of demographic change rather than sharp boundaries. In the case of spectral partitioning, this may be done by replacing the hard k-means subroutine with a Gaussian mixture model, or by using more advanced methods such as [154].

The primary limitation of these methods is their noninferential character. Though information theory is deeply intertwined with statistics, our methods use no explicit probabilistic model of spatial variation, making unavailable formal inferential procedures such as model selection. While the regionalization problem we solve bears resemblance to the problem of community-detection in annotated networks addressed by [145], that framework cannot accommodate spatial structure. An approach to segregation that supports both detailed spatial structure and formal inference in the context of segregation would be of considerable interest to both theorists and practitioners.

Chapter 4

Configuration Models of Random Hypergraphs

Graphs provide parsimonious mathematical descriptions of systems comprised of objects (nodes) and dyadic relationships (edges). When analyzing a given graph, a common task is to compare an observable of interest to its distribution under a suitably specified null model. A standard choice of null for dyadic networks is the class of *configuration models* [23, 37, 138, 76]. Configuration models preserve the degree sequence of the graph, which counts the number of edges incident to each node. These counts are natural first-order statistics of the graph, which are known to constrain many macroscopic graph properties [151]. Preserving these counts gives a natural null model constraint: properties observed in data that are not present in a configuration model require explanation in terms of higher-order graph structures.

In many systems of contemporary interest, groups of arbitrary size may interact simultaneously. Examples include social contact networks [195, 133]; scholarly and professional collaboration networks; [147, 17, 79, 168]; digital communications [118]; classifications on patents [217]; and many more [25]. In the past decades, the dominant approach to these systems has been to represent these networks dyadically, allowing the analyst to apply standard techniques of dyadic network science, including the configuration model. Recent work, however, has highlighted limitations of the dyadic paradigm in modeling of polyadic systems, in theory [183] and in applica-

tion domains including neuroscience [87], ecology [94], computational social science [202, 26] among others [25]. Scholarly collaboration networks provide a simple example of these limitations. Suppose that A, B, and C are collaborators in the same field. A three-author paper by A, B, and C is naturally represented by a polyadic edge of the form $e = (A, B, C)$. However, the standard dyadic representation is as *three* edges $e_1 = (A, B)$, $e_2 = (B, C)$, and $e_3 = (A, C)$. Note that this is exactly the same representation as would be obtained if each pair of authors wrote a two-author paper. The same dyadic representation thus fails to distinguish a world containing one three-author paper from a world containing three two-author papers. These worlds differ in the number of papers produced; the inferred productivity of the scholars; and likely in the content of the papers as well, since author lists are strong indicators of subject matter. In cases where these differences are relevant to our study questions, we can expect dyadic projection to yield misleading results.

The importance of polyadic interactions calls into question the use of the dyadic configuration model in such systems. It is therefore desirable to construct random models for polyadic data that inherit the useful properties of the dyadic configuration model. In this article, we construct two such models on suitably chosen spaces of hypergraphs, and demonstrate their utility for polyadic network data science. Along the way, we argue for two principle theses. First, the choice between dyadic and polyadic null models can determine the qualitative findings of standard network analyses. Second, the use of polyadic nulls allows the analyst to measure and test rich measures of polyadic structure, thereby expanding the network-scientific toolbox.

Outline

We begin in Section 4.1 with a survey of the landscape of null models for relational data, including the dyadic configuration model, random hypergraphs, and random simplicial complexes. In Section 4.2, we define stub- and vertex-labeled configuration models of random hypergraphs. Practical application of these models requires a sampling scheme, which we provide in Section 4.3. We turn to a triplet of illustrative applications in Section 4.4. We first consider triadic closure, showing that

some networks that would be considered clustered in comparison to dyadic nulls are significantly *less* clustered than the corresponding hypergraph nulls. We then turn to degree-assortativity, where hypergraph data representations allow us to define novel measures and conduct null hypothesis tests. Finally, we introduce a novel measure of correlation between polyadic edges, which can be tested against either the full configuration model or analytic approximations. We close in Section 4.5 with a summary of our findings and suggestions for future development.

4.1 Graphs, Hypergraphs, and Simplicial Complexes

Random graph null modeling has a rich history; see [76] for a review. In this section, we take a rapid tour through some of the most important results in configuration-type models for graphs and their generalizations. We begin with a brief review of the configuration model for dyadic graphs.

Definition 10 (Graph)**.** *A* graph $G = (V, E)$ *consists of a finite set* V *of nodes or vertices and a multiset* E*. An element of* E *is an unordered pair* $e = (u, v)$ *of nodes, also called an* edge*. We assume that both sets* V *and* E *are endowed with an (arbitrary) order. An edge of the form* (u, u) *is called a* self-loop*. Two distinct edges* e_1 *and* e_2 *are* parallel *if they are equal as sets.*

Let $n = |V|$ and $m = |E|$ be fixed. We denote by $\mathcal{G}^{\circlearrowleft}$ the set of all graphs on n nodes, and by $\mathcal{G} \subset \mathcal{G}^{\circlearrowleft}$ the set of graphs on n nodes without self-loops. Parallel edges are permitted in $\mathcal{G}$. While it is indeed possible to define configuration models on $\mathcal{G}^{\circlearrowleft}$ [76, 153], we do not do so here. We rule-out self-loops because (a) their presence considerably complicates sampling algorithms and (b) most polyadic data sets do not possess a meaningful notion of self-interaction. We will therefore present most of our results for elements of $\mathcal{G}$ and its generalization to hypergraphs, only discussing $\mathcal{G}^{\circlearrowleft}$ below in the context of certain technical issues.

The degree sequence of a graph $G = (V, E)$ is the vector $\mathbf{d} \in \mathbb{Z}^n$ defined compo-

nentwise as

$$d_v = \sum_{e \in E} \mathbb{I}(v \in e) \ . \tag{4.1}$$

A configuration model is a probability distribution on the set $\mathcal{G}_{\mathbf{d}} = \{G \in \mathcal{G} : \deg(G) = \mathbf{d}\}$ of graphs with degree sequence $\mathbf{d}$. There are two closely-related model variants which should be distinguished [76]. On its first introduction [37], the configuration model was defined mechanistically through a "stub-matching" algorithm. To perform stub-matching, we place d_v labeled half-edges (or "stubs") into an urn for each node v. We draw half-edges two at a time, with each draw producing an edge. A stub-labeled graph "remembers" which labeled stubs were drawn to form each edge.

Definition 11 (Stub-Labeled Graphs)**.** *For a fixed node set V and degree sequence $\mathbf{d}$, define the multiset*

$$\Sigma_{\mathbf{d}} = \biguplus_{v \in V} \{v_1, \ldots, v_{d_v}\} \ ,$$

where $\uplus$ denotes multiset union. The copies $v_1, \ldots, v_{d_v}$ are called stubs *of node v. A* stub-labeled graph *$S = (V, E)$ consists of the node set V and an edge set E which partitions $\Sigma_{\mathbf{d}}$ into unordered pairs. Each element of E has the form (u_i, v_j) for some nodes u and v and stub indices i and j. An edge of the form (v_i, v_j) is called a* self-loop*.*

Let $\mathcal{S}^{\circlearrowleft}$ be the set of stub-labeled graphs, and $\mathcal{S} \subset \mathcal{S}^{\circlearrowleft}$ the set without self-loops. Technically speaking, one should remember that the set $\mathcal{S}^{\circlearrowleft}$ of stub-labeled graphs is not a subset of the set $\mathcal{G}^{\circlearrowleft}$ of graphs, since the objects in the edge-set are of different logical types. The same is true of the sets $\mathcal{S}$ and $\mathcal{G}$. These technical considerations will also apply when we generalize to hypergraphs below, but will not present any major practical issues.

There is a natural surjection $g : \mathcal{S}^{\circlearrowleft} \to \mathcal{G}^{\circlearrowleft}$. If $S \in \mathcal{S}^{\circlearrowleft}$, $g(S) \in \mathcal{G}^{\circlearrowleft}$ is the graph obtained by erasing stub-labels: each stub v_i in S is recorded as an unlabeled copy of v in $g(S)$. The stub-labeled graph S and vertex-labeled graph $g(S)$ are topologically

identical, differing only in the presence of stub-labels in S. We use the notation $A = g^{-1}(G)$ to refer to the preimage $A \subseteq \mathcal{S}$ of $G \subseteq \mathcal{G}$ by g. We emphasize that g is not a bijection, and the symbol g^{-1} should not be interpreted as an inverse of g. We define $\mathcal{S}_{\mathbf{d}}^{\circlearrowleft}$ to be $g^{-1}\left(\mathcal{G}_{\mathbf{d}}^{\circlearrowleft}\right)$. Note that an edge $\tilde{e} \in S$ is a self-loop if and only if $e \in g(S)$ is. Because of this, $\mathcal{S} = g^{-1}(\mathcal{G})$. It is therefore natural to define $\mathcal{S}_{\mathbf{d}} = g^{-1}(\mathcal{G}_{\mathbf{d}})$. Fix $\mathbf{d} \in \mathbb{Z}_+^n$.

Definition 12 (Vertex-Labeled Dyadic Configuration Model [76])**.** *The* vertex-labeled configuration model *with degree sequence* $\mathbf{d}$ *is the uniform distribution* $\eta_{\mathbf{d}}$ *on* $\mathcal{G}_{\mathbf{d}}$*.*

Definition 13 (Stub-Labeled Dyadic Configuration Model [76])**.** *Let* $\lambda_{\mathbf{d}}$ *be the uniform distribution on* $\mathcal{S}_{\mathbf{d}}$*. The* stub-labeled configuration model *with degree sequence* $\mathbf{d}$ *is the distribution* $\mu_{\mathbf{d}} = \lambda_{\mathbf{d}} \circ g^{-1}$*.*

Here and below, the binary operator $\circ$ denotes composition of functions: $(\lambda_{\mathbf{d}} \circ g^{-1})(G) = \lambda_{\mathbf{d}}(g^{-1}(G))$.

In our formalism, the stub-labeled configuration model is not a distribution over the space of stub-labeled graphs $\mathcal{S}_{\mathbf{d}}$. Rather, it is the pushforward of such a distribution to the space $\mathcal{G}_{\mathbf{d}}$ of graphs. Intuitively, the vertex-labeled configuration model assigns the same probability to each graph with degree sequence $\mathbf{d}$, while the stub-labeled model weights these graphs according to their likelihood of being realized via stub-matching. One of the key insights of [37], since generalized by works such as [138, 11], is that these two models are related. Let $\mathcal{G}_{\text{simple}}$ be the set of *simple graphs*, which contain neither self-loops nor parallel edges. Then, $\mu_{\mathbf{d}}(G|G \in \mathcal{G}_{\text{simple}}) = \eta_{\mathbf{d}}(G|G \in \mathcal{G}_{\text{simple}})$. Furthermore, when the degree sequence is sampled from a fixed distribution with finite second moment, $\mu_{\mathbf{d}}(G \in \mathcal{G}_{\text{simple}})$ is bounded away from zero as n grows large (see, e.g. [11]), implying that repeated sampling from $\mu_{\mathbf{d}}$ will produce a simple graph in a number of repetitions that is asymptotically constant with respect to n. As a result, in the "large, sparse regime," it is possible to sample from the stub-labeled configuration model until a simple graph is obtained, which will then be distributed according to the vertex-labeled model. This relationship is extremely convenient, enabling asymptotic analytic expressions

for many quantities of theoretical and practical interest [151].

This close relationship between models is likely the reason why the distinction between them has often been elided in applied network science. Recently, however, the authors of [76] pointed out that, in many data sets, these two models are not interchangeable. It is important to distinguish them when the data may possess multi-edges or self-loops and the edge density is relatively high. The first condition is important because stub- and vertex-labeled models agree only on the subspace of simple graphs, not the full space of multigraphs. The second condition locates us away from the large, sparse regime and implies that parallel edges will occur under stub-matching with non-negligible probability.

From a modeling perspective, the choice of vertex- or stub-labeling must depend on domain-specific reasoning about counterfactual comparisons. Roughly, stub-labeling should be used when, for a fixed graph $G \in \mathcal{G}$, the elements of the set $g^{-1}(G) \subset \mathcal{S}$ have distinct identities in the context of the application domain. This corresponds to asking whether permutations of stubs lead to meaningfully different counterfactual data sets. In contrast, when stub-permutations are either nonsensical or are considered to leave the observed data unchanged, vertex-labeling is to be preferred. For example, in [76], the authors argue that vertex-labeled nulls are most appropriate for studying a collaboration network of computational geometers. Their reason is that stubs in this case correspond to an author's participation in a paper. It is nonsensical to say that A's first collaboration with B is B's second collaboration with A, and therefore stub-labeling is inappropriate.

Configuration models and their variants have played a fundamental role in the development of modern network science. The seminal paper by Molloy and Reed [137] has, according to Google Scholar, been cited at least 2,000 times since its publication, and over 800 times since 2015. How can we extend these models for application to polyadic data sets? A direct approach, taken in early studies such as [147], is to compute the *projected (dyadic) graph*. The projected graph represents each k-adic interaction as a k-clique, which contains an edge between each of the possible $\binom{k}{2}$ pairs of nodes (Figure 4-1). The resulting dyadic graph may then be randomized according

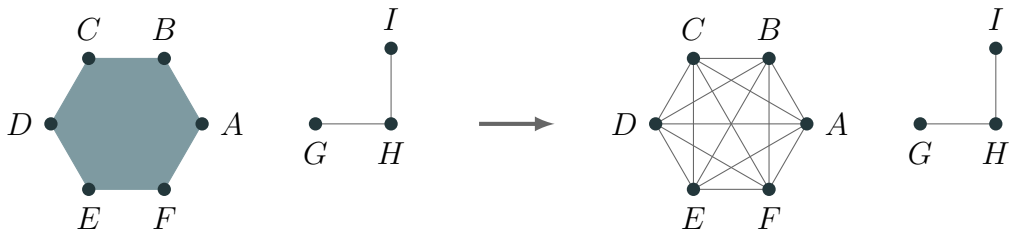


Figure 4-1: A synthetic coauthorship network with $n = 9$ nodes. On the left, the network represented as a hypergraph with 3 hyperedges. On the right, the projected graph with 17 dyadic edges.
.

to vertex- or stub-labeled dyadic configuration models. Projecting, however, can have unintended and occasionally counterintuitive consequences. First and most clearly, all properties which depend explicitly on the presence of higher-dimensional interactions are lost. Second, other observables such as node degrees and edge multiplicities may be transformed in undesirable ways; for example, a single interaction between six agents becomes 15 pairwise interactions after projection. As consequence, each of the six agents involved in a single 6-adic interaction are dyadically represented as nodes of degree 5. Third, and most subtly, projecting transforms the null space for downstream hypothesis-testing in ways that may not be intended. For example, projecting the network in Figure 4-1 prior to randomization implicitly chooses a null space of counterfactuals consisting of 17 two-author papers. This may be undesirable, especially when the null is viewed as a candidate data generating process. Given that the data possesses higher-order interactions, a null model that is by construction unable to produce such interactions may be implausible as a generator of relevant counterfactuals.

Random Hypergraphs

Considerations such as these motivate the development of dedicated null models for polyadic data. Such models enable the analyst to delay or omit dyadic projection when conducting null-hypothesis testing. We now make a brief survey of efforts to define configuration-type models for polyadic data. Hypergraphs provide the most

general context for such models. Hypergraphs are straightforward generalizations of graphs in which each edge is permitted to have an arbitrary number of nodes.

Definition 14 (Hypergraph)**.** *A* hypergraph $H = (V, E)$ *consists of a node set* V *and an edge set* $E = \{\Delta_j\}_{j=1}^{m}$ *which is a multiset of multisets of* V*. Each subset is called a* hyperedge*,* edge*, or, in some contexts, a* simplex*. Two hyperedges are* parallel *if they are equal as multisets. A hyperedge is* degenerate *if it contains two copies of the same node.*

Degenerate hyperedges generalize the notion of self-loops in dyadic graphs. We denote by $\mathcal{H}^{\circlearrowleft}$ the set of all hypergraphs and by $\mathcal{H}$ the set of all hypergraphs without degenerate edges. As before, parallel edges are permitted in $\mathcal{H}$. We continue to let $n = |V|$ and $m = |E|$.

Extant literature provides several approaches to defining null distributions on hypergraphs. One of the earliest approaches [151] takes a somewhat indirect route through bipartite graphs. A bipartite graph contains nodes of two classes, with connections permitted only between nodes of differing classes. To construct a bipartite graph B from a hypergraph H, one can construct a layer of nodes in B corresponding to the nodes V of H, and a second layer in B corresponding to the edges E of H. A node v is linked to an edge-node e iff $v \in e$ in H. We can now apply dyadic configuration models to randomize B, before recovering a hypergraph by projecting B onto its node layer. This approach is natural and convenient, but is only able to recover a generalization of the stub-labeled configuration model to hypergraphs. Generalizing the vertex-labeled model requires more complex tools which are not gracefully expressed in the bipartite formalism. We go into greater detail on this connection when discussing sampling methods in Section 4.3.

A more direct approach is to define a null distribution directly over $\mathcal{H}$. In [84], the authors define an analog of the stub-labeled configuration model over the set of hypergraphs in which all edges have three nodes, in the service of studying a tripartite tagging network on an online platform. Somewhat more general models have been formulated for the purposes of community-detection in hypergraphs via

modularity maximization, which requires the specification of a suitable null. In [119], the authors develop a degree-preserving randomization via a "corrected adjacency matrix," which may then be used for modularity maximization on the projected dyadic graph. In [111], the authors explicitly generalize the model of Chung and Lu [51], which preserves degrees in expectation, to non-uniform hypergraphs.

One subspecies of hypergraph has received additional attention. A *simplicial complex* is a hypergraph with additional structure imposed by a subset-inclusion relation: if $\Delta \in E$, then $\Gamma \in E$ for all $\Gamma \subseteq \Delta$. Simplicial complexes are attractive tools in studying topological aspects of discrete data [44], since the inclusion condition enables often-dramatic data compression while preserving topological features of interest. Configuration models of simplicial complexes provide one route for conducting null hypothesis tests of such features. The model of [58] achieves analytic tractability by restricting to simplicial complexes with maximal hyperedges of uniform dimension. The authors [218] allow heterogeneous dimensions but sacrifice analytic tractability, instead applying Markov Chain Monte Carlo to sample from the space. In applying any of these models, it is important to remember that subset-inclusion is strong property suited only to certain data-scientific contexts. Particular problems arise when edges possess the semantics of interaction, such as in collaboration networks. Suppose that authors A, B, and C jointly coauthor a paper. Using hypergraphs, we would represent this collaboration via an edge (A, B, C). In the setting of simplicial complexes, on the other hand, subset inclusion would also require us to include the edges (A, B), (B, C), (A, C), (A), (B), and (C). This may be undesirable, since we are not guaranteed that B and C, say, wrote a two-author paper. While simplicial complex modeling may be useful in carefully-selected application areas, in other cases we may require more flexible configuration models defined on more general spaces of polyadic data structures. We now formulate two such models.

4.2 Two Hypergraph Configuration Models

We now construct two configuration models for general hypergraphs. Our models generalize the stub- and vertex-labeled dyadic configuration models described in the previous section [76].

We use Greek letters to denote random edges of H, and English letters to denote nonrandom tuples of nodes. For example, the statement $\Delta = R$ describes the event that a random edge Δ has fixed location $R = (u_1, u_2, u_3, \ldots)$. Let $\binom{R}{\ell}$ denote the set of all subsets of R of size ℓ. Let $\mathbb{I}$ give the indicator function of its argument. We define the *degree sequence* $\mathbf{d} \in \mathbb{Z}_+^n$ and *dimension sequence* $\mathbf{k} \in \mathbb{Z}_+^m$ of a hypergraph H componentwise by

$$d_v = \sum_{e \in E} \mathbb{I}(v \in e) \quad \text{and} \quad k_e = \sum_{v \in V} \mathbb{I}(v \in e) \, .$$

Let $\mathcal{H}^{\circlearrowleft}_{\mathbf{d},\mathbf{k}}$ and $\mathcal{H}_{\mathbf{d},\mathbf{k}}$ denote the sets of hypergraphs with the specified degree and edge dimension sequences with and without degenerate hyperedges, respectively. We say that the sequences $\mathbf{d}$ and $\mathbf{k}$ are *configurable* if $\mathcal{H}_{\mathbf{d},\mathbf{k}} \neq \varnothing$.

Definition 15 (Vertex-Labeled Hypergraph Configuration Model)**.** *The* vertex-labeled configuration model $\eta_{\mathbf{d},\mathbf{k}}$ *is the uniform distribution on* $\mathcal{H}_{\mathbf{d},\mathbf{k}}$.

The stub-labeled configuration model is defined similarly as in the dyadic case.

Definition 16 (Stub-Labeled Hypergraphs)**.** *Let*

$$\Sigma = \biguplus_{v \in N} \{v_1, \ldots, v_{d_v}\}$$

be a multiset of stubs. *For each v, d_v copies of v appear in Σ. A* stub-labeled hypergraph S *has as its edge set E a partition of Σ in which each edge contains at most one stub for each node.*

The map g extends naturally to the space of hypergraphs. Let $\mathcal{S}$ be the set of stub-labeled hypergraphs. Then, if $S \in \mathcal{S}$, $g(S) \in \mathcal{H}$ is the hypergraph obtained by

erasing stub-labels: each stub v_i in S is recorded as an unlabeled copy of v in $g(S)$. Define $\mathcal{S}_{\mathbf{d},\mathbf{k}} = g^{-1}(\mathcal{H}_{\mathbf{d},\mathbf{k}})$

Definition 17 (Stub-Labeled Hypergraph Configuration Model)**.** *Let* $\lambda_{\mathbf{d},\mathbf{k}}$ *be the uniform distribution on* $\mathcal{S}_{\mathbf{d},\mathbf{k}}$*. Then, the* stub-labeled configuration model *is the distribution* $\mu_{\mathbf{d},\mathbf{k}} = \lambda_{\mathbf{d},\mathbf{k}} \circ g^{-1}$*.*

We have now defined two hypergraph configuration models, generalizing the vertex- and stub-labeled models of [76]. The vertex-labeled configuration model is the entropy-maximizing distribution on $\mathcal{H}_{\mathbf{d},\mathbf{k}}$ in the case that the identities of stubs are not meaningful, while the stub-labeled configuration model is the entropy-maximizing distribution when these identities are meaningful. The same considerations discussed in [76] (and briefly in the previous section) apply to the question of when to apply which null model.

Figure 4-2 illustrates some of the considerations in play. At left, we show two stub-labeled hypergraphs S_1 and S_2, represented as bipartite graphs. The associated hypergraph is $g(S_1) = g(S_2) = \{(u, v, w), (u, v, w)\}$, which contains two parallel edges of dimension 3. The stub-labeling is reflected as labels on the bipartite edges. Stub-labeled randomization treats each of S_1 and S_2 as distinct objects in sample space, while vertex-labeled randomization treats them as alternative representations of the same object. Because of this, stub-labeling should generally only be chosen when the distinct arrangements of stubs can be given valid interpretations in domain context. As the authors of [76] note, such cases are rare, and vertex-labeling is usually preferred. We further discuss the connection between hypergraphs and bipartite graphs in Section 4.3.1.

4.3 Sampling

Stub-matching is a classical method for sampling from the stub-labeled dyadic configuration model [37], and extends naturally to the case of random hypergraphs.

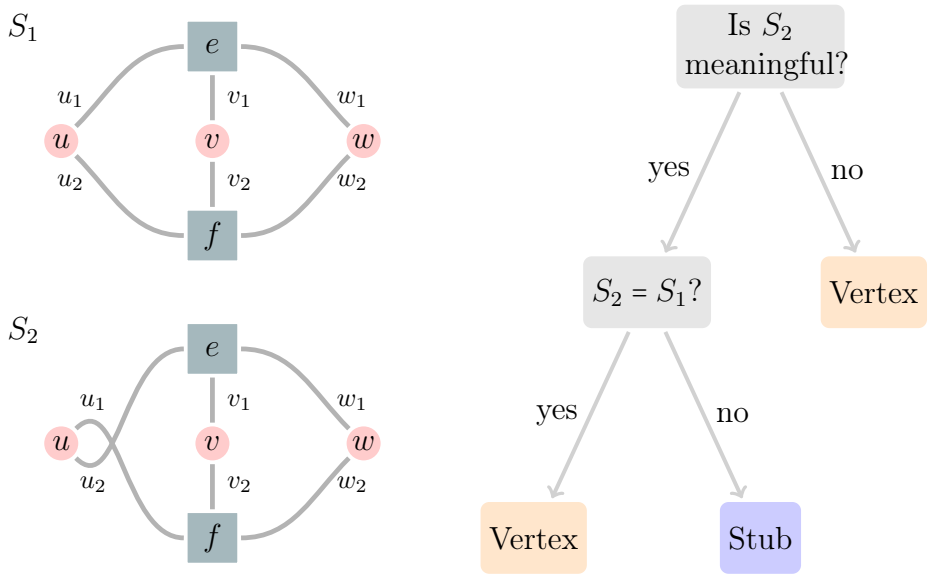


Figure 4-2: Choosing between stub- and vertex-labeled models.
(left): Bipartite representations of two stub-labeled hypergraphs S_1 and S_2 which correspond to the same hypergraph: $g(S_1) = g(S_2)$. The squares e and f denote edges; circles u, v, and w denote nodes, and an edge (u, e) denotes that $u \in e$. (right): A schematic flow chart for choosing stub- or vertex-labeled randomization depending on the interpretation of S_1 and S_2 in domain context.

Pseudcode for sampling from $\mu_{\mathbf{d},\mathbf{k}}$ via stub-matching is provided by Algorithm 3.

Algorithm 3: Hypergraph Stub-Matching

Input: Configurable $\mathbf{d} \in \mathbb{Z}_+^n$ and $\mathbf{k} \in \mathbb{Z}_+^m$.

Initialization: $j \leftarrow 1$, $S \leftarrow \varnothing$, $\Sigma \leftarrow \biguplus_{v \in V} \{v_1, \ldots, v_{d_v}\}$

for $j = 1, \ldots, m$ **do**

  $R \leftarrow \binom{W}{k_j}$

  $W \leftarrow W \setminus R$

  $S \leftarrow S \cup \{R\}$

Output: S

Since any stub-labeled graph S is as likely as any other under Algorithm 3, the output, conditioned on nondegeneracy, is distributed according to $\mu_{\mathbf{d},\mathbf{k}}$. As in the dyadic setting, there is nonzero probability for the output of stub-matching to produce a degenerate hypergraph. This probability will generally be large in the presence of highly heterogeneous node degrees – a common phenomenon in empirical data. Many iterations of Algorithm 3 may therefore be necessary in order to generate a single valid sample. Because of this, pure stub-matching is often not a practical method for generating random hypergraphs. That said, the stub-matching algorithm is often useful in proofs involving $\mu_{\mathbf{d},\mathbf{k}}$.

For practical sampling, we consider a Markov Chain Monte Carlo (MCMC) approach, in which we use successive, small alterations to the edge-set E in order to systematically explore the space $\mathcal{H}_{\mathbf{d},\mathbf{k}}$. Our scheme extends to hypergraphs the broad class of edge-swap MCMC samplers, which have also been formulated for marginal-constrained binary matrices [205, 13]; degree-regular [206, 135, 109] and degree-heterogeneous [45, 197, 32, 61] simple graphs; degree-constrained graphs [76]; bipartite graphs with degree constraints [112]; and graphs with prescribed degree correlations [6].

Definition 18 (Pairwise Reshuffle)**.** *Let $S \in \mathcal{S}_{\mathbf{d},\mathbf{k}}$, and $\Delta, \Gamma \in S$. A* pairwise reshuffle *$b(\Delta, \Gamma | S)$ of Δ and Γ is a sample from the conditional distribution $\mu(\cdot | E \setminus \{\Delta, \Gamma\})$. Depending on context, we will regard a pairwise reshuffle as either a random map on stub-labeled hypergraphs or on pairs of hyperedges.*

Lemma 1. *Let $S \in \mathcal{S}$. Let $b(\Delta, \Gamma | H) = (\Delta', \Gamma')$ be a pairwise reshuffle which results in $S' \in \mathcal{S}$. Then,*

1. *The degree and dimension sequences are preserved:* $\deg(S) = \deg(S')$ *and* $\dim(S) = \dim(S')$.

2. *We have* $\Delta' \cap \Gamma' = \Delta \cap \Gamma$.

3. *Any given realization of b occurs with probability*

$$q_\mu(\Delta, \Gamma) = 2^{-|\Delta \cap \Gamma|} \binom{|\Delta| + |\Gamma| - 2|\Delta \cap \Gamma|}{|\Delta| - |\Delta \cap \Gamma|}^{-1}. \tag{4.2}$$

Proof. A pairwise reshuffle may be performed via the following sequence, which is an alternative description of the final two iterations (conditioning on nondegeneracy) of Algorithm 3.

1. Delete Δ and Γ from E.

2. Construct Δ' and Γ' as (initially empty) node sets.

3. For each node $v \in \Delta \cap \Gamma$, add a v-stub to both Δ' and Γ'.

4. From the remaining stubs, select $|\Delta \setminus \Gamma|$ stubs uniformly at random and add them to Δ'. Add the remainder to Γ'.

5. Add Δ' and Γ' to E.

Each node begins with the same number of edges as it started, so degrees are preserved. Next, by construction, $|\Delta'| = |\Delta \cap \Gamma| + |\Delta - \Gamma| = |\Delta|$, and similarly $|\Gamma'| = |\Gamma|$. The edge dimension sequence is thus also preserved.

Finally, by construction, step 2 above preserves the intersection $\Delta \cap \Gamma$. There are $2^{|\Delta \cap \Gamma|}$ ways to assign stubs to this intersection. There are a total of $|\Delta| + |\Gamma| - 2|\Delta \cap \Gamma|$ remaining stubs, and of these one must choose $|\Delta| - |\Delta \cap \Gamma|$ to be placed in Δ. We infer that any given pairwise reshuffle is realized with probability given by (4.2), as was to be shown. □

We now define a transition kernel of a first-order Markov chain on the space $\mathcal{S}_{\mathbf{d},\mathbf{k}}$. Write $S \sim_{\Delta,\Gamma} S'$ if there exists a pairwise shuffle b such that $b(\Delta,\Gamma|S) = S'$. Note that, since each element of each edge has a distinct label in $\mathcal{S}_{\mathbf{d},\mathbf{k}}$, for any S and S' there is at most one pair (Δ,Γ) such that $S \sim_{\Delta,\Gamma} S'$. If no such pair exists, we write $S \not\sim S'$. Then, let

$$\tilde{p}_\mu(S'|S) = \begin{cases} \binom{m}{2}^{-1} q_\mu(\Delta,\Gamma) & S \sim_{\Delta,\Gamma} S' \\ 0 & S \not\sim S' \,, \end{cases} \tag{4.3}$$

where $q_\mu(\Delta,\Gamma)$ is the number of distinct possible shuffles realizable from Δ and Γ; an explicit expression is given in the SI. To sample from $\tilde{p}_\mu(\cdot|S)$, it suffices to sample two uniformly random edges from E and perform a reshuffle. The prefactor $\binom{m}{2}^{-1}$ gives the probability that any two given edges are chosen.

The sequence $\{S_t\}$ is Markovian by construction. The following lemma and its corollary ensure that the sequence $\{H_t\} = \{g(S_t)\}$ is also a Markov chain.

Lemma 2. *Let $H, H' \in \mathcal{H}$. Suppose that $S_1, S_2 \in g^{-1}(H)$ and $S_1', S_2' \in g^{-1}(H')$. Then, $\tilde{p}_\mu(S_1'|S_1) = \tilde{p}_\mu(S_2'|S_2)$.*

Proof. The objects S_1 and S_2 may each be considered arbitrary stub-labelings of part-edges in H. Similarly, S_1' and S_2' are each arbitrary labelings of part-edges in H'. However, by (4.3), $\tilde{p}_\mu(\cdot|S)$ depends only on the sizes of edges and their intersections in H, not their labels. □

Corrolary 6. *The process $\{H_t\} = \{g(S_t)\}$ on $\mathcal{H}_{\mathbf{d},\mathbf{k}}$ is a Markov chain.*

Proof. Markovianity of $\{H_t\}$ follows from Lemma 2. Indeed, we can construct H_t mechanistically from H_{t-1} by choosing $S_{t-1} \in g^{-1}(H_{t-1})$, setting $S_t \sim \tilde{p}_\mu(\cdot|S_{t-1})$, and then letting $H_t = g(S_{t-1})$. Lemma 2 ensures that the distribution of H_t depends only on H_{t-1}, and not on the choices of S_{t-1} and S_t. □

Theorem 3. *The Markov chain $\{S_t\}$ on $\mathcal{S}_{\mathbf{d},\mathbf{k}}$ defined by the kernel $\tilde{p}_\mu$ is irreducible and reversible with respect to $\lambda_{\mathbf{d},\mathbf{k}}$, the uniform distribution on $\mathcal{S}_{\mathbf{d},\mathbf{k}}$. If in addition at*

least two entries of $\mathbf{k}$ *are two or larger,* $\{S_t\}$ *is also aperiodic. In this case,* $\lambda_{\mathbf{d},\mathbf{k}}$ *is the equilibrium distribution of* $\{S_t\}$*. Furthermore,* $\mu_{\mathbf{d},\mathbf{k}}$ *is the equilibrium distribution of the process* $\{H_t\} = \{g(S_t)\}$*.*

Proof. We will first show reversibility with respect to $\lambda_{\mathbf{d},\mathbf{k}}$. Fix S. Let $S \sim_{\Delta,\Gamma} S'$ and $S' \sim_{\Delta',\Gamma'} S$. In this case, we have $\Delta', \Gamma' = b(\Delta, \Gamma | S)$. Then,

$$\tilde{p}_\mu(S'|S) = \binom{m}{2}^{-1} q_\mu(\Delta, \Gamma) = \binom{m}{2}^{-1} q_\mu(\Delta', \Gamma') = \tilde{p}_\mu(S|S') \ ,$$

as required. The second equality follows from Lemma 1, since $q(\Delta, \Gamma)$ depends only on $|\Delta|$, $|\Gamma|$, and $|\Delta \cap \Gamma|$.

Our proof approach for irreducibility generalizes that of [76]. We need to construct a path of nonzero probability between two arbitrary elements S_1 and S_2 of $\mathcal{S}$. Let E_1 and E_2 be the edge-sets of S_1 and S_2, respectively. We first describe a procedure for generating a new stub-labeled hypergraph S_3 such that $|E_2 \setminus E_3| < |E_2 \setminus E_1|$. Since $E_1 \neq E_2$ and $|E_1| = |E_2|$, we may pick $\Delta = \{\delta_1, \ldots, \delta_\ell\} \in E_2 \setminus E_1$. Note that, since $\Delta \notin E_1$ and the edge dimension sequences must agree, there exists an edge $\Psi \in E_1 \setminus E_2$ such that $|\Psi| = |\Delta| = \ell$. Now, for each i, since $\Delta \notin E_1$, δ_i belongs to a different edge (call it Γ_i) in E_1. Note that we may have $\Gamma_i = \Gamma_{i'}$ in case δ_i and $\delta_{i'}$ belong to the same hyperedge in E_1. Suppose we have $j \leq \ell$ such edges. Since δ_i is a stub, δ_i can belong to only one edge in each hypergraph, and therefore $\Gamma_k \notin E_2$ for each $k = 1, \ldots, j$. For each $k = 1, \ldots, j$, let $\left(\Psi_k, \Gamma'_k\right) = b_k(\Psi_{k-1}, \Gamma_k)$, where b_k assigns all elements of the set $\Delta \cap (\Psi_{k-1} \cup \Gamma_{k-1})$ to Ψ_k and uniformly distributes the remainder. Since $\Delta \subseteq \left(\bigcup_{k=1}^{j} \Gamma_k\right)$ by construction, by the end of this procedure we have $\Psi_j = \Delta$. Call the resulting stub-labeled hypergraph S_3 with edge set E_3. Since we have only modified the edges $\{\Gamma_k\}$ and Ψ, which are elements of $E_1 \setminus E_2$, we have not added any edges to the set $E_1 \setminus E_2$, but we have removed one, namely Ψ. We therefore have $|E_2 \setminus E_3| < |E_2 \setminus E_1|$, as desired. Applying this procedure inductively, we obtain a path of nonzero probability between S_1 and S_2, proving irreducibility.

To prove aperiodicity, we will construct supported cycles of length 2 and 3 in $\mathcal{S}$. Since the lengths of these cycles are relatively prime, aperiodicity will follow.

To construct a cycle of length 2, pick two edges Δ and Γ and any valid reshuffle $b : (\Delta, \Gamma) \mapsto (\Delta', \Gamma')$. Then, $b^{-1} : (\Delta', \Gamma') \mapsto (\Delta, \Gamma)$ is also a valid reshuffle, and the sequence (b, b^{-1}) of transitions constitutes a supported cycle through $\mathcal{S}$ of length 2. To construct a cycle of length 3, choose two edges Δ and Γ which each contain two or more nodes, writing $\Delta = \{\delta_1, \delta_2, \ldots\}$ and $\Gamma = \{\gamma_1, \gamma_2, \ldots\}$. This is always possible by hypothesis. Then, the following sequence of pairwise reshuffles constitutes a cycle of length 3:

$$\begin{aligned}\{\delta_1, \delta_2, \ldots\}, \{\gamma_1, \gamma_2, \ldots\} &\mapsto \{\gamma_1, \delta_2, \ldots\}, \{\delta_1, \gamma_2, \ldots\} \\ &\mapsto \{\gamma_2, \delta_2, \ldots\}, \{\delta_1, \gamma_1, \ldots\} \\ &\mapsto \{\delta_1, \delta_2, \ldots\}, \{\gamma_1, \gamma_2, \ldots\} \, .\end{aligned}$$

We have shown reversibility, irreducibility, and aperiodicity, completing the proof. □

A small modification enables sampling from the vertex-labeled model $\eta_{\mathbf{d},\mathbf{k}}$. Let m_Δ give the number of edges parallel to edge Δ in hypergraph H, including Δ itself. Define

$$a_\eta(S'|S) = \begin{cases} \frac{2^{|\Delta \cap \Gamma|}}{m_\Delta m_\Gamma} & S \sim_{\Delta,\Gamma} S' \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

Theorem 4. *Let $\tilde{p}_\eta(S'|S) = a(S'|S)\tilde{p}_\mu(S'|S)$. Let $\{S_t\}$ be the Markov chain generated by $\tilde{p}_\eta$. Then, the process $\{H_t\} = \{g(S_t)\}$ is a Markov chain. Furthermore, $\{H_t\}$ is irreducible and reversible with respect to $\eta_{\mathbf{d},\mathbf{k}}$. If in addition $\mathbf{k}$ has at least two entries larger than 2, $\{H_t\}$ is aperiodic. In this case, $\eta_{\mathbf{d},\mathbf{k}}$ is the equilibrium distribution of $\{H_t\}$.*

Proof. Markovianity of $\{H_t\}$ follows from the same argument as Corollary 6. Irreducibility and aperiodicity follow from Theorem 3, since the state space $\mathcal{H}$ is a partition of $\mathcal{S}$ into equivalence classes induced by g. It remains to demonstrate reversibility with respect to $\eta_{\mathbf{d},\mathbf{k}}$. Let p_η be the transition kernel of H_t. Fix H and H'.

Fix $S \in g^{-1}(H)$ and $S^* \in g^{-1}(H')$. Then, we can write

$$\begin{aligned} p_\eta(H'|H) &= \sum_{S' \in g^{-1}(H')} \tilde{p}_\eta(S'|S) \\ &= \sum_{S' \in g^{-1}(H')} a(S'|S) \tilde{p}_\mu(S'|S) \\ &= a(S^*|S) \sum_{S' \in g^{-1}(H')} \tilde{p}_\mu(S'|S) \, . \end{aligned}$$

The expressions appearing in this calculation are independent of the specific choices of S or S^* following the same argument as in the proof of Lemma 2. We now evaluate the sum in the third line. The summand is nonzero if and only if $S \sim_{\Delta,\Gamma} S'$, in which case its value depends only on $|\Delta|$, $|\Gamma|$, and $|\Delta \cap \Gamma|$. We therefore count terms. There are $2^{|\Delta \cap \Gamma|}$ ways to arrange the intersection of Δ and Γ in $\mathcal{S}$, and $m_\Delta m_\Gamma$ ways to choose two edges parallel to Δ and Γ to reshuffle, all of which generate a distinct element of $g^{-1}(H')$. The sum therefore possesses precisely $a(S^*|S)^{-1}$ terms. We find that $p_\eta(H'|H) = \tilde{p}_\mu(S'|S)$ for any $S \in g^{-1}(H)$ and $S' \in g^{-1}(H')$. Reversibility of p_η thus follows from reversibility of $\tilde{p}_\mu$. □

Algorithm 5 supplies pseudocode for sampling from the stub- and vertex-labeled

hypergraph configuration models.

Algorithm 4: Markov Chain Monte Carlo for hypergraph configuration models

Input: $\mathbf{d}$, $\mathbf{k}$, target distribution $\nu \in \{\mu_{\mathbf{d},\mathbf{k}}, \eta_{\mathbf{d},\mathbf{k}}\}$, initial hypergraph $H_0 \in \mathcal{H}_{\mathbf{d},\mathbf{k}}$, sample interval $h \in \mathbb{Z}_+$, desired sample size $s \in \mathbb{Z}_+$.

Initialization: $t \leftarrow 0$, $H \leftarrow H_0$

for $t = 1, 2, \ldots, sh$ **do**

  sample (Δ, Γ) uniformly at random from $\binom{E_t}{2}$

  $H' = b(\Delta, \Gamma | H_t)$

  if Uniform$([0,1]) \leq a_\nu(H'|H)$ **then**

    $H_t \leftarrow H'$

  else

    $H_t \leftarrow H_{t-1}$

Output: $\{H_t$ such that $t|h\}$

Theorems 3 and 4 constitute a guarantee that, for sufficiently large sample intervals h, the hypergraphs sampled from algorithm 5 will be asymptotically i.i.d. according to the desired distribution. Unfortunately, we are unaware of any mixing-time bounds for this class of Markov chain. It is therefore possible in principle that the scaling in the mixing time as a function of system size is extremely poor, a result suggested by work on related classes of edge-swap Markov chains [92, 93, 70]. Our experience indicates, however, that sampling is possible for configuration models with many thousands of edges in practical time. For example, the `email-enron` used here contains $10,887$ edges. In the implemented code, it is possible to take 10^5 steps of stub-labeled MCMC in under four seconds on personal computing equipment, and of vertex-labeled MCMC in under thirty. The significantly larger timing for vertex-labeled MCMC is due to the incorporation of the acceptance probability, which implies that a successful transition occurs less frequently. In addition to potential code optimizations, many sampling tasks can also be parallelized, leading to shorter compute times when required.

4.3.1 Connections to Random Bipartite Graphs

As briefly mentioned in Section 4.1, a hypergraph $H = (V, E)$ corresponds in a natural way to a bipartite dyadic graph B. The graph B consists of a node set $V \cup E$. An edge (u, e) exists between $u \in V$ and $e \in E$ iff $u \in e$ (in H). In this setting, the degree of u (in H) is equal to its degree in B, and the dimension of e (in H) is similarly equal to its degree in B. Let h be the function that assigns to each hypergraph its associated bipartite graph. When both nodes and edges are uniquely labeled, h is a bijection. It follows that a probability measure ν on the space $\mathcal{B}_{\mathbf{d},\mathbf{k}}$ of bipartite graphs with node degrees $\mathbf{d}$ and $\mathbf{k}$ induces a probability measure $\nu \circ h^{-1}$ on $\mathcal{H}_{\mathbf{d},\mathbf{k}}$. Several extant papers (e.g. [151, 182]) use this equivalence to construct random models of polyadic data. While it is sometimes thought that bipartite randomization supplies a complete solution to null hypergraph sampling, we show in this section that the natural scope of the bipartite method is limited to stub-labeled models.

We first define a bipartite, dyadic, configuration model. We define $\nu_{\mathbf{d},\mathbf{k}}$ to be the measure on $\mathcal{B}_{\mathbf{d},\mathbf{k}}$ obtained by performing stub-matching with the node-set $V \cup E$, conditioned on the events that (a) all edges have the form (u, e) for $u \in V$ and $e \in E$, and (b) the bipartite graph is simple, without multi-edges or self-loops. Note that conditioning on the event that B is simple implies that the stub-labeled and vertex-labeled models are identical. The work of Kannan et al. [112] considers the problem of sampling from $\nu_{\mathbf{d},\mathbf{k}}$ via bipartite edge-swaps. Such a swap maps $(u, e), (v, f) \mapsto (u, f), (v, e)$. By construction, such a swap preserves $\mathbf{d}$ and $\mathbf{k}$. The authors show that a Markov chain which performs successive, random bipartite edge-swaps (while avoiding ones that would lead to a non-simple bipartite graph) is ergodic and therefore sufficient to sample from $\nu_{\mathbf{d},\mathbf{k}}$. Such a swap, which viewed in the space $\mathcal{H}_{\mathbf{d},\mathbf{k}}$ amounts to swapping the edge memberships of nodes u and v. Importantly, a sequence of such switches is special case of the pairwise reshuffle Markov chain on $\mathcal{S}_{\mathbf{d},\mathbf{k}}$. This implies following relationship:

Proposition 1. *The configuration model on simple bipartite graphs is equivalent to the stub-labeled hypergraph configuration model, in the sense that* $\mu_{\mathbf{d},\mathbf{k}} = \nu_{\mathbf{d},\mathbf{k}} \circ h^{-1}$.

Proposition 1 makes precise the primary sense in which bipartite randomization provides an approach to random hypergraph modeling. This is a convenient result, since a dyadic edge-swap Markov chain on B can be used to produce samples from $\mu_{\mathbf{d},\mathbf{k}}$. This equivalence may also be used to give alternative proofs of Theorem 3. However, as discussed in [76], many data sets in which we aim to apply null modeling are better represented by vertex-labeled null distributions. There is no obvious route for vertex-labeled sampling through bipartite random graphs. In particular, there is no analogue of Proposition 1 for this case. Thus, even though the work of [112] treats vertex-labeled sampling from $\mathcal{B}_{\mathbf{d},\mathbf{k}}$, this does not directly suffice for vertex-labeled sampling from $\mathcal{H}_{\mathbf{d},\mathbf{k}}$. The reason is that sampling from the vertex-labeled measure $\eta_{\mathbf{d},\mathbf{k}}$ requires adjusting for permutations of parallel hyperedges. When H contains multiple hyperedges of dimension three or greater, it is necessary to track multiple node-edge incidence relations in order to check when hyperedges are parallel.

It is possible to write down a version of Algorithm 5 for vertex-labeled sampling in which the fundamental data structure is a bipartite graph rather than a hypergraph. However, the result would not, to the author's knowledge, correspond to any standard random bipartite graph model. Expressing both models directly on the space $\mathcal{H}_{\mathbf{d},\mathbf{k}}$ of hypergraphs supports both conceptual clarity and a convenient formulation of MCMC for both stub- and vertex-labeled models. Incidentally, we note that this discussion constitutes another, separate setting in which adherence to dyadic methods can limit our data-analytical horizons. An exclusive focus on nulls realizable through bipartite methods obscures the possibility of vertex-labeled polyadic models.

4.4 Network Analysis with Random Hypergraphs

We now illustrate the application of hypergraph configuration models through three simple network analyses. We first study triadic closure in polyadic networks, finding that the use of polyadic nulls can generate different, interpretable study findings when compared to dyadic nulls. We then turn to degree-assortativity, defining and testing three distinct measures of association via polyadic data representations and

randomizations. Finally, we study the tendency of edges to intersect on multiple vertices in the `email-Enron` data set, finding using simulation and analytical methods that large intersections occur at much higher rates than would be expected by random chance. Collectively, these cases illustrate the use of polyadic methods to define and analyze richer measures of network structure, and the use of polyadic nulls in interpreting the results.

The data sets for case study were gathered, cleaned, and generously made public by the authors of [25]. In certain experiments, data were temporally filtered in order to reduce their size; these cases have been explicitly noted in the text and the filtering procedure described in Section 4.5.1. Importantly, in no case was the filtering operation motivated by the expense of Monte Carlo sampling; rather, the bottlenecks were standard, expensive computations such as triangle-counting in dyadic graphs.

4.4.1 Triadic Closure

Triadic closure refers to the phenomenon that, in many networks, if two nodes u and v interact with a third node w, then it is statistically likely that u and v also interact with each other. Studies such as [196, 147, 151] observed triadic closure in many empirical networks, and highlighted the fact that dyadic configuration models tend to be unable to reproduce this behavior. Traditionally, triadic closure is measured by a ratio of the number of triangles (closed cycles on three nodes) that are present in the graph, compared to the number of "wedges" (subgraphs on three nodes in which two edges are present).[1] Local and global variants of this ratio have been proposed. We follow the choice of [196] and work with the *average local clustering coefficient*. Let T_v denote the number of triangles incident on v, and W_v the number of wedges. Note that $W_v = \binom{d_v}{2}$. The average local clustering coefficient is

$$\bar{C} = \frac{1}{|N|} \sum_{v \in N} \frac{T_v}{W_v} \,. \tag{4.5}$$

[1] Recent measures have been developed for higher-order notions of clustering on larger subgraphs; see [216].

		Hypergraph		Projected	
	$\bar{C}$	Vertex	Stub	Vertex	Stub
`congress-bills`*	0.608	0.601(1)	0.622(2)	0.451(2)	0.611(1)
`coauth-MAG-Geology`*	0.8200	0.8196(7)	0.8186(7)	0.00035(3)	0.00035(3)
`email-Enron`	0.658	0.825(3)	0.808(4)	0.638(5)	0.797(3)
`email-Eu`*	0.540	0.569(4)	0.601(4)	0.398(4)	0.598(4)
`tags-ask-ubuntu`*	0.571	0.609(4)	0.631(5)	0.183(4)	0.499(6)
`threads-math-sx`*	0.293	0.435(3)	0.426(3)	0.041(1)	0.093(2)

Table 4.1: Hypothesis-tests for clustering coefficients of selected data sets. Average local clustering coefficients for selected data sets, compared to their expectations computed under vertex- and stub-labeling of hypergraph and projected graph models. Parentheses show standard deviations in the least-significant figure under the equilibrium distribution of each null model. Starred* data sets have been temporally filtered as described in Section 4.5.1.

It is direct to show that, in dyadic configuration models and under mild sparsity assumptions, $\bar{C}$ decays to zero as n grows large [150, 196].

The average local clustering coefficient $\bar{C}$ is a natively dyadic metric, in the sense that "wedges" and "triangles" are defined explicitly in terms of 2-edges. To compute $\bar{C}$ in polyadic data, it is therefore necessary to project a hypergraph down to a dyadic graph. In the context of hypothesis-testing, there is some subtlety involved in the choice of when to do this. One method is to project first and then randomize via a dyadic null model. This is the most common historical approach, used for example in [151]. Alternatively, one may randomize via polyadic nulls prior to projection. This approach has the effect of preserving clustering induced by polyadic edges, since an edge of dimension k contains $3\binom{k}{3}$ wedges and $3\binom{k}{3}$ ordered triangles.

Table 4.1 summarizes a sequence of experiments performed on two collaboration networks (top) and four communication networks (bottom). For each network, we computed the observed local clustering coefficient $\bar{C}$ on the unweighted projected graph. We choose the unweighted projected graph, rather than the weighted projected graph, in order to more closely match previous analyses (such as those of [151]), as well as to avoid ambiguities that arise in the definition of $\bar{C}$ in the presence of multi-

edges. We then compared the observed value to its null distribution under four randomizations. We first randomized using the vertex- and stub-labeled hypergraph configuration models, *prior* to projecting and measuring $\bar{C}$. These results are shown in the second and third columns. We then reversed the order, first computing the projected graph and randomizing via dyadic configuration models. The results are shown in the fourth and fifth columns.

Benchmarking against dyadic configuration models yields mixed results. Vertex-labeled configuration models conclude in all cases that the observed degree of clustering is significantly higher than would be expected by random chance. Stub-labeled benchmarking concludes that `congress-bills` and the two email data sets have significantly less clustering than expected, while the remainder have significantly more. The stub-labeled results should be approached with caution – for reasons discussed in detail in [76], the stub-labeled configuration model is a less-relevant comparison for these data sets than the vertex-labeled model.

Hypergraph randomization leads to different conclusions. First, the expected values of $\bar{C}$ under both hypergraph vertex- and stub-labeled nulls are much closer together than under dyadic nulls, indicating that the polyadic statistical test is much less sensitive than the dyadic test to the choice of vertex- and stub-labeling. Second, the vertex-labeled null separates the two collaboration networks from the four communication networks. These two data sets are only slightly more clustered than expectation under the vertex-labeled model. Chebyshev's inequality implies that, at 95% confidence, `congress-bills` would be considered "significantly more clustered" than its expectation under the vertex-labeled model, while the opposite conclusion would be reached under the stub-labeled model. On the other hand, `coauth-MAG-Geology` falls within two standard deviations of its expectation under each model; it would be necessary to inspect the full sampling distribution to conduct a significance test at 95% confidence in this case. In contrast, Chebyshev's inequality can be used to show that the four communication networks are all significantly *less* clustered than either vertex- or stub-labeled nulls would expect. Not only is there no clustering beyond that implied by the edge dimensions; triadic closure even appears to be inhibited in

these data sets.

From a purely statistical perspective, these examples highlight the importance of careful null model selection in hypothesis-testing for triadic closure. More physically, use of hypergraph nulls allows us in this case to distinguish data sets by their generative mechanisms. The communication networks are all less clustered than expected, while the collaboration networks are approximately as clustered as expected. This result is to some extent intuitive. Collaborations between many agents often have nontrivial coordination costs that scale with the number of agents involved. It may be easier to assemble and coordinate a set of overlapping groups than a single large collective. In such cases, one may expect to observe clustering near or above that expected at random, since overlaps between related groups would generate triangles. In contrast, in digital communications it is essentially effort-free to construct interactions between larger groups of agents. Examples include adding an email address to the "cc" field or introducing participants to thread on a forum. In such cases, triangles composed of distinct edges are energetically unnecessary, and may reflect redundant information flow. We therefore hypothesize that these systems have a tendency to absorb potential triangles into higher-dimensional interactions. This results in lower levels of clustering than would be expected under polyadic nulls. These considerations hint toward the importance of studying edge correlations via natively polyadic metrics as we do in Section 4.4.3.

4.4.2 Degree-Assortativity

A network is degree-assortative when nodes of similar degrees preferentially interact with each other. Early studies found that different categories of social, biological, and technological networks display different patterns of assortative mixing by degree [149, 148, 57]. Social networks, for example, are frequently measured to be degree-assortative. In this context, degree-assortativity is often taken to indicate a tendency for popular or productive agents to interact with each other.

We measure degree-assortativity in hypergraphs via a generalization of the standard Spearman rank assortativity coefficient to hypergraphs. Importantly, there are

multiple possible generalizations, each of which measures distinct structural information about degree correlations. Let $E_{\geq 2} = \{\Delta \in E : |\Delta| \geq 2\}$. Let $h : E_{\geq 2} \to N^2$ be a (possibly random) *choice function* that assigns to each edge Δ two distinct nodes $u, v \in \Delta$. Three possibilities of interest are:

$$h(\Delta) = (u, v) \sim \text{Uniform}\binom{\Delta}{2} \tag{Uniform}$$

$$h(\Delta) = (u, v) = \underset{w,w' \in \binom{\Delta}{2}}{\text{argmax}}\, d_w d_{w'} \tag{Top-2}$$

$$h(\Delta) = (u, v) = \left(\underset{w \in \Delta}{\text{argmax}}\, d_w,\ \underset{w \in \Delta}{\text{argmin}}\, d_w\right) \tag{Top-Bottom}$$

The Uniform choice function selects two distinct nodes at random. The Top-2 choice function selects the two distinct nodes in the edge with largest degree. The Top-Bottom choice function selects the nodes with largest and smallest degree.

Let $r : N \to \mathbb{R}$ be a ranking function on the node set; we will always take $r(u)$ to be the rank of node u by degree in the hypergraph. For fixed h, let $f : E \to \mathbb{R}^2$ be defined componentwise by $f_j(\Delta) = (r \circ h_j)(\Delta)$. Then, the *generalized Spearman assortativity coefficient* is the empirical correlation coefficient between $f_1(\Delta)$ and $f_2(\Delta)$:

$$\rho_h = \frac{\sigma^2\left(f_1(\Delta), f_2(\Delta)\right)}{\sqrt{\sigma^2\left(f_1(\Delta), f_1(\Delta)\right)\sigma^2\left(f_2(\Delta), f_2(\Delta)\right)}}, \tag{4.6}$$

where $\sigma^2(X, Y) = \langle XY \rangle - \langle X \rangle \langle Y \rangle$ and brackets express averages over pairs of edges in E.

In the case of dyadic graphs, the three choice functions above are trivially identical, since there is only one way to pick two nodes from an edge of size two. On polyadic data, however, the resulting Spearman coefficients capture usefully different classes of information. For example, in studying coauthorship networks, they may be used to test hypotheses such as the following:

1. **Generic Assortativity**: On a given paper, most coauthors will simultaneously be more or less prolific than average.

2. **Junior-Senior Assortativity**: The least profilic author on a paper will tend

to be relatively more prolific if the most prolific author is relatively more prolific.

3. **Senior-Senior Assortativity**: The two most prolific authors on a paper will tend to be simultaneously more or less prolific than average.

While the corresponding Spearman coefficients may in general be correlated, substantial variation manifests across study data sets. Figure 4-3 shows measurements and significance tests for one synthetic data set and the six empirical data sets studied in the previous section. The synthetic data consists of five copies of the hypergraph shown in Figure 4-1. For each data set, we compute the dyadic assortativity coefficient on the projected graph (first column), as well as each of the three polyadic assortativity coefficients defined above.

The synthetic data (first row) illustrates a stark case in which dyadic hypothesis-testing leads to a finding of statistically-significant assortativity, while polyadic hypothesis-testing finds statistically-significant *dis*assortativity. In each of the empirical data sets, the dyadic and polyadic tests show qualitative agreement. However, the polyadic tests highlight several features of the data missed by the dyadic tests. In the two email data sets, all four coefficients are positive and to the right of the null distributions, though projecting (first column) increases the significance of the coefficients relative to hypergraph randomization (second column). The two forum data sets (`threads-math-sx` and `tags-ask-ubuntu`) are disassortative when compared to vertex-labeled nulls. The fact that `tags-ask-ubuntu` is disassortative despite a positive uniform hypergraph Spearman coefficient speaks to the importance of carefully specified null hypothesis testing. Interestingly, the misspecified stub-labeled randomization would lead to the opposite finding. The coauthorship network `coauth-MAG-Geology` is highly assortative in all metrics – including the top-bottom measure, which is negative. The `congress-bills` data set is also assortative in all measures. Unlike the other data sets, the uniform hypergraph coefficient lies farther from the bulk of its null distribution than does the projected coefficient. We note in passing that, whereas the stub-labeled and vertex-labeled hypergraph distributions had similar expected rates of triadic closure table 4.1, their distributions of degree-assortativity coefficients vary

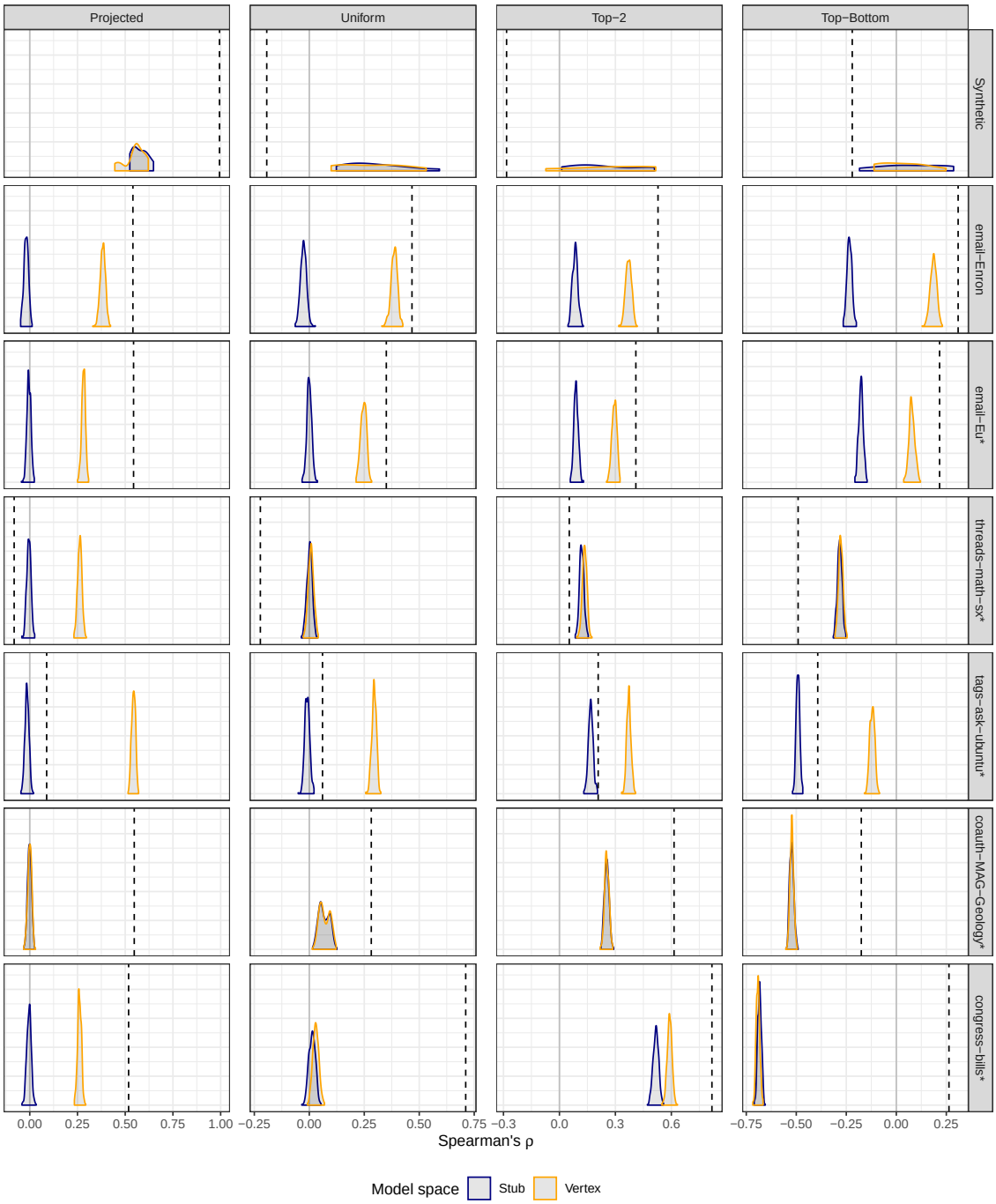


Figure 4-3: Significance tests of degree-assortativity in synthetic and empirical networks.

The synthetic data consists of five copies of the illustrative network shown in Figure 4-1. In each figure, the dashed line gives the observed Spearman correlation, and densities give the null distributions under vertex- and stub-labeled configuration models. In the first column only, the hypergraph was projected down to an unweighted dyadic graph prior to randomization. Starred* data sets have been temporally filtered as described in the SI.

substantially, and would in some cases lead to qualitatively different study conclusions.

When studying triadic closure, we saw how hypergraph null models could lead us to contextualize standard graph metrics differently. When studying assortativity, we gain even more. Use of hypergraph nulls allows us to forgo the dyadic projection operation, and thereby define rich polyadic assortativity measures. Hypergraph null models thus enable us to expand our network-analytic toolboxes by measuring and interpreting novel structural patterns in polyadic data.

4.4.3 Hyperedge Intersection Profiles

Let $\Delta, \Gamma \in H$. What is the size of their intersection? In the case of dyadic graphs, the intersection can have size at most two, when Δ is parallel to Γ. In hypergraphs intersections of arbitrary sizes may occur. The existence of large intersections in a data set may indicate the emergence of polyadic social ties between groups of agents, or interpretable event sequences such as email threads or series of related scholarly papers. Several recent papers [25, 160] have studied similar questions by considering the rate at which "holes" in the hypergraph tend to be "filled in" by higher-order interactions. We take a simpler approach, defining a measure which is both easily computed and amenable to analytical approximation.

Definition 19 (Intersection Profile)**.** *For fixed $k, \ell \in \mathbb{Z}_+$, the* conditional intersection profile *of a hypergraph $H \in \mathcal{H}$ is the distribution*

$$r_{k\ell}(j|H) = \langle \mathbb{I}(|\Delta \cap \Gamma| = j) \rangle_{k\ell} \, ,$$

where $\langle \cdot \rangle_{k\ell}$ denotes the empirical average over all hyperedges Δ of size k and Γ of size ℓ. The marginal intersection profile *is*

$$r(j|H) = \langle \mathbb{I}(|\Delta \cap \Gamma| = j) \rangle \, ,$$

with the average taken over all pairs of distinct edges in E.

Large values of $r_{k\ell}(j|H)$ indicates that edges of size k and ℓ frequently have intersections of size j in H. Empirical data sets may possess complex patterns of correlation between edges of various sizes. Evaluating whether an observed conditional or marginal intersection profile is noteworthy requires comparison to appropriately-chosen null models.

Figure 4-4 demonstrates the use of hypergraph configuration models to study the intersection profile of the `email-Enron` data set. In Figure 4-4(a), we compare the empirical average intersection size $\langle J \rangle_{k\ell} = \sum_{j=0}^{\infty} j r_{k\ell}(j|H)$ to its average $\langle \hat{J} \rangle_{k\ell}$ under the vertex-labeled configuration model. Higher values of the ratio $\frac{\langle J \rangle_{k\ell}}{\langle \hat{J} \rangle_{k\ell}}$ indicate the presence of denser intersections between edges of sizes k and ℓ. Notably, the empirical averages are not uniformly higher than the null model averages, even on the diagonal. There is apparent block structure, indicating that edges of certain sizes tend to correlate most strongly with certain other sizes. Edges of dimension 3 through 6 tend to interact strongly with each other, as do edges of dimension 7 and 8. However, edges in the smaller group interact more weakly with edges in the larger group than would be expected by chance. Further, more detailed study may be able to shed light on the groups of agents involved in these overlapping communications.

Figure 4-4(b) gives a global view of the data using the marginal intersection profile. The observed profile (points in Figure 4-4(b)) is nearly linear on semilog axes through $j = 6$, suggesting that the decay in the intersection size is roughly exponential. In order to evaluate whether this behavior indicates nonrandom clustering between edges, we again turn to hypergraph configuration models. The expectation $\hat{r}(j) = \mathbb{E}_{\nu}[r(j|H)]$ of the marginal intersection profile under a configuration model $\nu \in \{\mu_{\mathbf{d},\mathbf{k}}, \eta_{\mathbf{d},\mathbf{k}}\}$ measures the typical behavior of a comparable random hypergraph. The solid lines in Figure 4-4(b) give these expected profiles under both stub- and vertex-labeled models, which qualitatively agree. The observed data shows fewer intersections on single vertices than would be expected by chance. On the other hand, for $j \geq 3$, $r(j|H)$ exceeds $\hat{r}(j)$ by an order of magnitude or more, suggesting substantial higher-order correlation in the data. These results likely reflect the passing of multiple messages between the same sets of users.

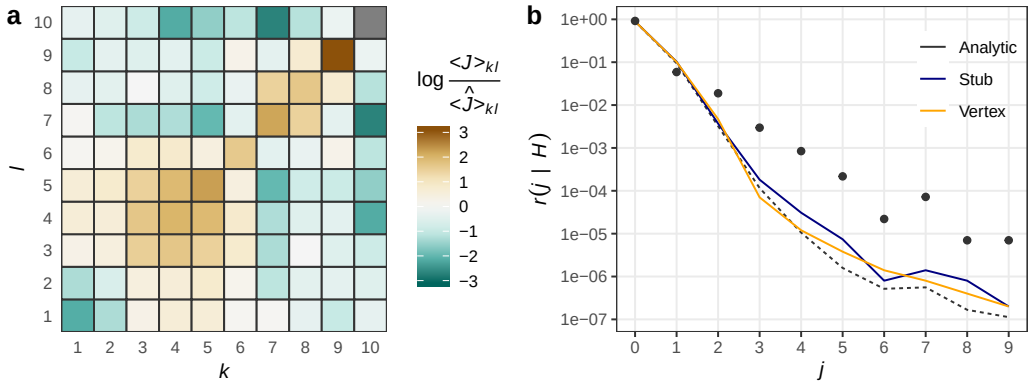


Figure 4-4: Analysis of intersection profiles in the `email-Enron` data set. (**a**): The average of the intersection size normalized by the expectation $\langle \hat{J} \rangle_{k\ell}$ under the vertex-labeled configuration model. Positive values indicate that the data displays larger intersections than expected under the configuration model for the corresponding values of k and ℓ. Colors are shown on a log scale. The missing value at $(k, \ell) = (10, 10)$ indicates that no nonempty intersections were observed between edges of these sizes in the Monte Carlo sampling runs. (**b**): Marginal intersection profile (points) of the empirical data, compared to null distributions under the stub- and vertex-labeled configuration models. The dashed gray line gives the analytic approximation of (4.7). Note the logarithmic vertical axis.

Some data sets may be too large to practically estimate $\hat{r}(j)$ by Monte Carlo methods. In such cases, it is possible to approximate $\hat{r}(j)$ under the stub-labeled configuration model analytically, using the following asymptotic result.

Theorem 5. *Fix ℓ, k, and j. Let $\mathbf{D} \in \mathbb{Z}_+^n$ be a vector of i.i.d. copies of positive, discrete random variable $D \in \mathbb{Z}_+$ such that $D \leq d_{\max}$ with probability one for some $d_{\max}$. Let $\mathbf{K} \in \mathbb{Z}_+^m$ be any vector of edge dimensions configurable with $\mathbf{D}$. Let $H \sim \mu_{\mathbf{D},\mathbf{K}}$, and let Δ and Γ be uniformly random edges of H. Then, with high probability (w.h.p.) as n grows large,*

$$\hat{r}_{k\ell}(j) = (1 + O(n^{-1}))j! \binom{k}{j} \binom{\ell}{j} \left(\frac{1}{n} \frac{\mathbb{E}[D^2] - \mathbb{E}[D]}{\mathbb{E}[D]^2} \right)^j . \tag{4.7}$$

Proof. Let $\langle d \rangle = \frac{1}{n} \sum_{u \in N} d_u$ denote the empirical mean degree of a given degree sequence $\mathbf{d}$. Assume without loss of generality that $\Delta = \{\delta_1, \ldots, \delta_k\}$ and $\Gamma = \{\gamma_1, \ldots, \gamma_\ell\}$ are the first two hyperedges formed by Algorithm 3, conditioned on nondegeneracy.

There are $\binom{k}{j}$ ways to choose the j elements of Δ contained in $\Delta \cap \Gamma$, and similarly $\binom{\ell}{j}$ ways to choose the elements of Γ. There are then $j!$ ways to place these two sets in bijective correspondence. Define the event $A = \{\delta_h = \gamma_h,\ h = 1, \ldots, j\}$. Then, $\hat{r}_{k\ell}(j) = j!\binom{k}{j}\binom{k}{\ell}\mu_{\mathbf{D},\mathbf{K}}(A)$. To compute $\mu_{\mathbf{D},\mathbf{K}}(A)$, we may explicitly enumerate

$$\mu_{\mathbf{D},\mathbf{K}}(A) = \sum_{u \in N} \frac{d_u}{n\langle d \rangle} \frac{d_u - 1}{n\langle d \rangle - 1} \left[\sum_{v \in N \setminus \{u\}} \frac{d_v}{n\langle d \rangle - d_u} \frac{d_v - 1}{n\langle d \rangle - d_u - 1} \left[\sum_{w \in N \setminus \{u,v\}} \cdots \right] \right],$$

with a total of j sums appearing. In each summation, the first factor gives the probability that $\delta_1 = u$ and the second that $\gamma_1 = u$. Consider the innermost summation, which may be written

$$S_R = \sum_{z \in N \setminus R} \frac{d_z}{n\langle d \rangle - \sum_{y \in R} d_y} \frac{d_z - 1}{n\langle d \rangle - \sum_{y \in R} d_y - 1} \tag{4.8}$$

for a set R of size $j - 1$. Since $D \leq d_{\max}$ with probability one by hypothesis, we may employ Chebyshev's inequality to find that $(n\langle d \rangle)^{-1} \sum_{y \in R} d_y = O(n^{-1})$ w.h.p. We may therefore w.h.p. expand both factors within (4.8), obtaining the expression

$$\sum_{z \in N \setminus R} \frac{d_z(d_z - 1)}{n^2 \langle d \rangle^2} \left(1 + O(n^{-1})\right) .$$

Using Chebyshev's inequality again, we also obtain asymptotic behavior on the other expressions appearing above. $\langle d \rangle = (1 + O(n^{-1}))\,\mathbb{E}[D]$ and $\sum_{z \in N \setminus R} \frac{d_z(d_z-1)}{n} = (1 + O(n^{-1}))\,(\mathbb{E}[D^2] - \mathbb{E}[D])$, both w.h.p. We have therefore shown that

$$S_R = \left(1 + O\left(n^{-1}\right)\right) \left(\frac{1}{n} \frac{\mathbb{E}[D^2] - \mathbb{E}[D]}{\mathbb{E}[D]^2} \right) \tag{4.9}$$

w.h.p. This argument may be repeated inductively for each of the remaining $j-1$ sums, each of which contributes the same factor appearing in (4.9), proving the theorem. □

Figure 4-4 shows the resulting approximation for $\hat{r}_{k\ell}$ as a dashed line, finding excellent qualitative agreement. This approximation may be used to study intersection profiles in data sets of arbitrary size. Figure 4-5 shows the use of this approxima-

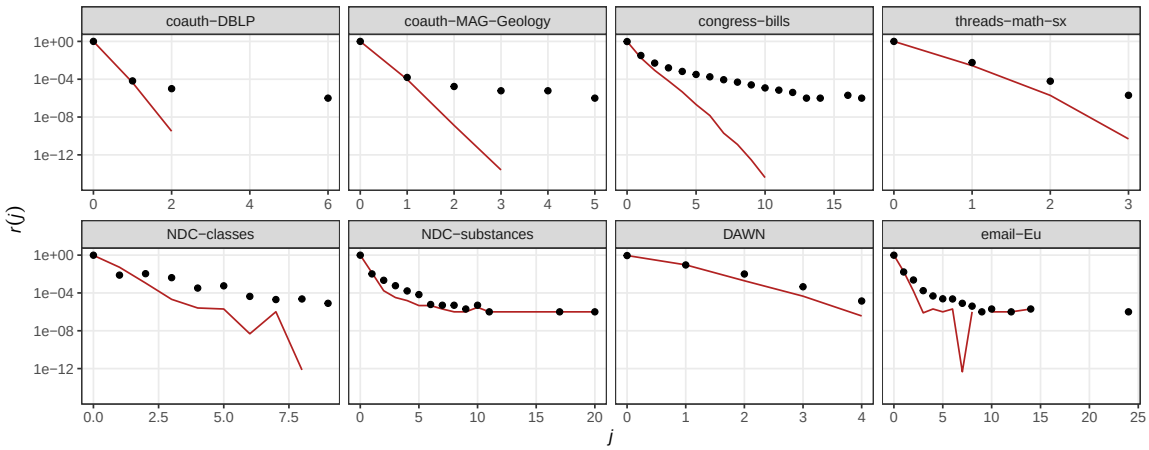


Figure 4-5: Empirical and analytically-approximate intersection profiles for large polyadic data sets.
Points give the observed intersection profile for six large polyadic data sets. The solid line gives the null intersection profile of Theorem 5. In this visualization, full data sets were used – the temporal filtering described in Section 4.5.1 was not performed.

tion to study intersection profiles in hypergraph data sets of arbitrary size. The top set of panels shows four data sets in which the approximate null intersection profile consistently underestimates the rates of large intersections by several orders of magnitude, clearly indicating the presence of correlation structure over and above what would be expected under hypergraph randomization. The lower panels show four additional data sets in which the approximate null profile more closely-approximates the observed data.

4.5 Discussion

Configuration models of random hypergraphs preserve the first moments of the data – the degree and edge-dimension sequences – while remaining maximally ignorant about additional data structure. These models extend the widely-used configuration models for dyadic graphs, and serve as natural null models for polyadic network data analysis. We have demonstrated how to define, sample from, analyze, and apply these models. We have seen that the choice between nulls can greatly impact the qualitative findings of studies of empirical polyadic data. The analyst faced with

such a choice must therefore carefully consider whether dyadic simplification will lead to data representations and null spaces that are relevant for their application area. Second, employing polyadic nulls often allows the analyst to define novel measures that can illuminate higher-order structure in data. We have illustrated this with extended assortativity measures and intersection profiles, but many more extensions are possible. We hope that the introduction of statistically-grounded hypergraph nulls will encourage analysts to design, measure, and carefully interpret many novel measures of polyadic network structure.

There are several directions of future work on configuration models of random hypergraphs. Beginning with theory, many classical asymptotic results on dyadic configuration models invite generalization. These include probabilistic characterization of component sizes; cycles and parallel edges; and the diameter of the connected component in various regimes. We also highlight two applications of potential interest. The first is motif analysis. A network motif is a subgraph that appears with higher-than-expected frequency in a given network [136], relative to a given null model. Considering the explicit dependence of this definition on the null, we conjecture that motif-discovery algorithms based on polyadic nulls may highlight importantly distinct structure when compared to dyadic nulls. A second promising application is in hypergraph clustering and community detection. A recent paper [111] offers a definition of modularity — a common quality function for network partitioning — based on a polyadic generalization of the Chung-Lu model [51]. In this case, the modularity of a given partition may be computed analytically. The same calculations used to prove Theorem 5 can also be used to show that the stub-labeled configuration model will give an asymptotically equivalent expression. However, for the large class of data sets more appropriately modeled by vertex-labeled nulls, other methods may be necessary. We anticipate that pursuing these tasks will pose interesting theoretical and computational challenges.

Funding

This work was supported by the National Science Foundation Graduate Research Fellowship under award number 1122374.

Acknowledgments

I am grateful to Patrick Jaillet for helpful discussions from which this work benefited substantially.

Software

A `hypergraph` class, written in Python 3.5, is available at

`https://github.com/PhilChodrow/hypergraph`.

This class includes implementations of Monte Carlo sampling for both stub- and vertex-labeled configuration models, as well as a simple tutorial illustrating the use of the software.

Additionally, a repo illustrating the analysis pipeline – including data acquisition, batch computations, and visualization scripts – is available at

`https://github.com/PhilChodrow/hypergraph_analysis`.

This repository has considerably more moving parts, and is recommended only to those aiming to reproduce the results of this paper. Those who wish only to conduct novel analyses with hypergraph configuration models should refer to the first repository.

4.5.1 Data

The data sets used in this chapter were prepared by the authors of [25] and accessed from `https://www.cs.cornell.edu/~arb/data/`. Some data sets have been filtered

to exclude edges prior to a temporal threshold τ in order to promote practical compute times on triangle counting and mixing of vertex-labeled models in projected graph spaces. Notably, in no cases was sampling from hypergraph configuration models the computational bottleneck. Thresholds were chosen to construct data with edge sets of approximate size $m \approx 10^4$, but are otherwise arbitrary. Temporal data subsets were used in the generation of Table 4.1 and fig. 4-3. Table 4.2 gives the node and edge counts of both the original data and the data after temporal subsetting when applicable.

	Original			Filtered	
	n	m	τ	n	m
`email-Enron`	143	10,886	–	–	–
`email-Eu`	1,006	235,264	1.105×10^9	817	32,117
`congress-bills` [78, 79]	1,719	260,852	7.315×10^5	537	6,661
`coauth-MAG-Geology` [191]	1,261,130	1,591,167	2017	73,436	23,434
`threads-math-sx`	201,864	719,793	2.19×10^{12}	11,880	22,786
`tags-ask-ubuntu`	200,975	192,948	2.6×10^{12}	2,120	19,338

Table 4.2: Summary of data preparation. When τ is given, the filtered data set consists in all edges that occurred after time τ.

Chapter 5

Moments of Uniform Random Multigraphs with Fixed Degree Sequences

We study the expected adjacency matrix of a uniformly random multigraph with fixed degree sequence $\mathbf{d} \in \mathbf{Z}_+^n$. This matrix arises in a variety of analyses of networked data sets, including modularity-maximization and mean-field theories of spreading processes. Its structure is well-understood for large, sparse, simple graphs: the expected number of edges between nodes i and j is roughly $\frac{d_i d_j}{\sum_\ell d_\ell}$. Many network data sets are neither large, sparse, nor simple, and in these cases the standard approximation no longer applies. We derive a novel estimator using a dynamical approach: the estimator emerges from the stationarity conditions of a class of Markov Chain Monte Carlo algorithms for graph sampling. We derive error bounds for this estimator, and provide an efficient scheme with which to compute it. We test the estimator on synthetic and empirical degree sequences, finding that it enjoys relative error against ground truth a full order of magnitude smaller than the standard approximation. We then compare modularity maximization techniques using both the standard and novel estimator, finding that the qualitative structure of the optimization landscape depends significantly on the estimator choice. Our results emphasize the importance of using carefully specified random graph models in data scientific applications.

This chapter is a reproduction of the manuscript [48], currently submitted and under revision.

5.1 Introduction

The language of graphs offers a standard formalism for representing systems of interrelated objects or agents. Simple graphs model agents connected by a single, usually static, relation, such as acquaintanceship, proximity, or similarity. In many data sets, however, agents are linked by multiple, discrete interactions. Two agents in a contact network may be in spatial proximity multiple times in the study period. Two agents in a communication network may exchange many emails over the course of a week. In an academic collaboration network, the same two authors may be jointly involved in tens or even hundreds of papers. In such cases, it is natural to draw a distinct edge between agents for each interaction event. Doing so results in a multigraph, in which any two nodes may be linked by an arbitrary, nonnegative, integer-valued number of edges.

A fundamental tool in network data science is null model comparison, which allows the analyst to evaluate whether a feature observed in a given network is surprising when compared to benchmark expectations. We therefore often compare observed networks against random graph null models – probability distributions over graphs. An especially common class of null models is obtained by fixing the degree sequence $\mathbf{d}$ of the observed network, which encodes the number of interactions for each node. The degree sequence is known to constrain many of a network's macroscopic properties [151]. The least informative (or entropy-maximizing) distribution so obtained is the uniform distribution on the space of graphs with the specified degree sequence. The same construction goes through for multigraphs. When studying interaction networks, the corresponding random graph is the uniform distribution $\eta_{\mathbf{d}}$ on the set $\mathcal{G}_{\mathbf{d}}$ of multigraphs with degree sequence $\mathbf{d}$.

In many applications, a set of complete samples from $\eta_{\mathbf{d}}$ is not required – only some selected moments. An especially important set of moments is summarized by

the expected adjacency matrix. We therefore consider the following question: if $\mathbf{W}$ is the (random) adjacency matrix of multigraph $G \sim \eta_{\mathbf{d}}$, what is the value of the expected adjacency matrix $\mathbf{\Omega} \triangleq \mathbb{E}[\mathbf{W}]$? The entry ω_{ij} of $\mathbf{\Omega}$ gives the expected number of edges between nodes i and j. These moments have several important applications in network science. Among these is community-detection via modularity-maximization [143], which in many formulations includes a term for the expected number of edges between nodes under a suitably specified null model. Despite its simplicity and relevance for applications, this problem has received relatively little mathematical attention.

Before surveying existing approaches to the estimation of $\mathbf{\Omega}$, we fix some notation. Let $\mathcal{G}_{\mathbf{d}}$ refer to the set of multigraphs without self-loops with degree sequence $\mathbf{d} \in \mathbf{Z}_+^n$. From a modeling perspective, the exclusion of self-loops reflects an assumption that agents do not meaningfully interact with themselves. An element $G \in \mathcal{G}_{\mathbf{d}}$ has a fixed number n of nodes and $m = \frac{1}{2}\sum_i d_i$ of edges. We use bold uppercase symbols to denote matrices, bold lowercase symbols to denote vectors, and standard symbols to denote scalars. We do not notationally distinguish deterministic and random objects, instead relying on their associated definitions. We use Greek letters to denote expectations of random objects. An estimator of a quantity, either deterministic or stochastic, is distinguished by a hat. For example, $\mathbf{\Omega} = \mathbb{E}[\mathbf{W}]$ is the expectation of $\mathbf{W}$. An estimator of $\mathbf{\Omega}$, either deterministic or stochastic, may be written $\hat{\mathbf{\Omega}}$.

One approach to estimating $\mathbf{\Omega}$ is Monte Carlo sampling. We sample s independent and identically distributed samples $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(s)} \sim \eta_{\mathbf{d}}$, and construct the estimator

$$\hat{\mathbf{\Omega}}^{\mathrm{mc}}(\mathbf{d}) \triangleq \frac{1}{s}\sum_{\ell=1}^{s} \mathbf{W}^{(\ell)} \,. \tag{5.1}$$

The estimator $\hat{\mathbf{\Omega}}^{\mathrm{mc}}$ is a random function of $\mathbf{d}$, parameterized by the sample size s. The Strong Law of Large Numbers (SLLN) ensures that $\hat{\mathbf{\Omega}}^{\mathrm{mc}} \to \mathbf{\Omega}$ almost surely as the number of samples s grows large. Stronger results are possible: since each entry $\hat{\omega}_{ij}^{\mathrm{mc}}$ is bounded, the variance of $\hat{\omega}_{ij}^{\mathrm{mc}}$ is finite and we can apply the Central Limit Theorem to provide quantitative bounds on the convergence rate. This attractive picture

is marred by a severe computational inconvenience: the size and complex combinatorial structure of $\mathcal{G}_d$ makes exact sampling intractable. Markov Chain Monte Carlo (MCMC) methods [76] are therefore required. MCMC introduces a new complication: for any finite number of iterations, the samples produced will always be statistically dependent, and may therefore over-represent some regions of $\mathcal{G}_{\mathbf{d}}$ and under represent others. This dependence breaks the guarantees provided by the SLLN or Central Limit Theorem. Control over the *mixing time* of the sampler is in principle sufficient to ameliorate this issue; however, there are few known mixing time bounds on MCMC samplers of distributions on $\mathcal{G}_{\mathbf{d}}$. Available upper bounds on the mixing times [92, 93, 70] are too large for guarantees in many practical computations, and there are heuristic reasons to believe that there are limits on our ability to improve these bounds.

An alternative estimator $\hat{\mathbf{\Omega}}^0$, extremely common in the network science literature, is defined entrywise by a simple formula:

$$\hat{\omega}^0_{ij}(\mathbf{d}) = f_{ij}(\mathbf{d}) \triangleq \begin{cases} \frac{d_i d_j}{2m} & i \neq j \\ 0 & i = j \ . \end{cases} \tag{5.2}$$

The function f_{ij} plays an important role throughout this article. Unlike $\hat{\mathbf{\Omega}}^{\text{mc}}$, $\hat{\mathbf{\Omega}}^0$ is a deterministic function of $\mathbf{d}$ that is essentially free to compute. The functional form of $f_{ij}(\mathbf{d})$ can be derived in multiple ways. For example, it is the expected edge density between distinct nodes i and j in the model of Chung and Lu [52, 51], which preserves $\mathbf{d}$ in expectation rather than deterministically. We will therefore refer to (5.2) as the "CL estimate" after Chung and Lu, though we emphasize that these authors did not use this expression as an estimator for any of the models we consider here, and indeed restricted their attention to graphs without parallel edges. The estimator $\hat{\mathbf{\Omega}}^0$ was also derived heuristically by Newman and Girvan when they introduced modularity maximization as a method for community detection in networks [146, 143]. In their derivation, we approximate the number of edges between i and j as follows. Node i has d_i edges. Each of these edges must connect to one of the $n-1$ other nodes. A

"random edge" is attached to node j with probability roughly $\frac{d_j}{2m-d_i}$. Assuming that $d_i \ll 2m$ yields $\hat{\omega}_{ij}^0$ as an approximation. It is important to note that this heuristic argument does not formalize any probability measure over a set of graphs. Thus, although $\hat{\mathbf{\Omega}}^0$ is sometimes described as the expectation of a "random graph with fixed degree sequence," this is not exactly true for any common models except that of Chung and Lu, in which degrees are fixed only in expectation. In particular, $\hat{\mathbf{\Omega}}^0$ possesses no guarantees related to its performance as an estimator for the uniform model $\eta_{\mathbf{d}}$, the most literal mathematical operationalization of the phrase "random graph with fixed degree sequence." As we will see this performance can indeed be quite poor on data sets with high edge densities.

In this article, we construct an estimator of $\mathbf{\Omega}$ for dense multigraphs that is both scalable and accurate. By treating an MCMC sampler for $\eta_{\mathbf{d}}$ as a stochastic dynamical system whose state space is $\mathcal{G}_{\mathbf{d}}$, we derive stationarity conditions describing the desired moments. As we will show, there exists a vector $\boldsymbol{\beta} \in \mathbb{R}_+^n$ such that χ_{ij}, the probability that $w_{ij} \geq 1$, is given by

$$\chi_{ij} \triangleq \eta_{\mathbf{d}}(w_{ij} \geq 1) \approx \frac{\beta_i \beta_j}{\sum_i \beta_i} = f_{ij}(\boldsymbol{\beta})$$

for all $i \neq j$. The function f_{ij} in this approximation is the same as that which appears in the definition of the CL estimator in (5.2). Furthermore, the entries of $\mathbf{\Omega}$ are given approximately by

$$\omega_{ij} \approx \frac{\chi_{ij}}{1-\chi_{ij}} \, .$$

Taken to together, these two formulae provide a method for computing an estimate of $\mathbf{\Omega}$ given knowledge of the vector $\boldsymbol{\beta}$. We construct an estimator $\hat{\boldsymbol{\beta}}$ of this vector by solving the system of n nonlinear equations

$$\sum_j \frac{f_{ij}(\boldsymbol{\beta})}{1-f_{ij}(\boldsymbol{\beta})} = d_i \, , \quad i = 1, \ldots, n \, .$$

We show that the solution to this equation, provided it exists, is unique within the

realm of interpretable sequences $\boldsymbol{\beta}$ subject to mild regularity conditions, and that this solution can be found efficiently by a simple, iterative algorithm. From $\hat{\boldsymbol{\beta}}$ we construct an estimator $\hat{\boldsymbol{\Omega}}^{1}$ of $\boldsymbol{\Omega}$. As we show, this estimator is both easier to compute than $\hat{\boldsymbol{\Omega}}^{\mathrm{mc}}$ and much more accurate than $\hat{\boldsymbol{\Omega}}^{0}$. Furthermore, we can view the Chung-Lu estimator $\hat{\boldsymbol{\Omega}}^{0}$ as an approximation of $\hat{\boldsymbol{\Omega}}^{1}$, obtained from the latter via a sequence of two linear approximations.

5.1.1 Outline

In Section 5.2, we review two important null multigraph models – the configuration model and the uniform model – as well as a unified MCMC algorithm for sampling from each. The analysis of this algorithm forms the heart of our derivation of the estimate $\hat{\boldsymbol{\Omega}}^{1}$ in Section 5.3. This estimator depends on the unknown vector $\boldsymbol{\beta}$, which must be learned from $\mathbf{d}$. We offer a simple scheme for doing so in Section 5.4, including a qualified uniqueness guarantee on the resulting estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$; a description of its structure; and a numerical scheme for computing it efficiently. In Section 5.5 we turn to experiments. We first study the behavior of our methods on two synthetic data sets, including a bootstrap-style test of the conjecture underlying our error-bounds. We then check the accuracy of $\hat{\boldsymbol{\Omega}}^{1}$ on a subset of a high school contact network. Whereas $\hat{\boldsymbol{\Omega}}^{0}$ is significantly biased on this data set, $\hat{\boldsymbol{\Omega}}^{1}$ is nearly unbiased and decreases the mean relative error of the estimate by an order of magnitude. In our final experiment, we study the behavior of modularity maximization when the standard null expectation $\hat{\boldsymbol{\Omega}}^{0}$ is replaced by $\hat{\boldsymbol{\Omega}}^{1}$. We find that the behavior of a multiway spectral algorithm [221] depends strongly on both the choice of null expectation and the data set under study. We close in Section 5.6 with a discussion and suggestions for future work.

5.2 Random Graphs with Fixed Degree Sequences

Our interest will focus on the uniform model $\eta_{\mathbf{d}}$, but it will be useful draw comparisons to the somewhat more commonly-used configuration model [37].

Definition 20 (Configurations)**.** *For a fixed node set N and degree sequence $\mathbf{d} \in \mathbb{Z}_+^n$, let*

$$\Sigma_{\mathbf{d}} = \biguplus_{i=1}^{n} \{i_1, \ldots, i_{d_i}\} \ ,$$

where $\uplus$ denotes multiset union. Thus, $\Sigma_{\mathbf{d}}$ contains d_i labeled copies of each node i. The copies $i_1, \ldots, i_{d_i}$ are called stubs *of node i. A* configuration *$C = (N, E)$ consists of the node set N and an edge set E which partitions $\Sigma_{\mathbf{d}}$ into unordered pairs. An edge in E of the form $\{i_k, i_\ell\}$ is called a* self-loop*. The process of forming C from $\Sigma_{\mathbf{d}}$ is often called* stub-matching*.*

Let $\mathcal{C}_{\mathbf{d}} \subset \Sigma_{\mathbf{d}}$ be the set of all configurations with degree sequence $\mathbf{d}$ that do not include any self-loops. There is a natural surjection $g : \mathcal{C}_{\mathbf{d}} \to \mathcal{G}_{\mathbf{d}}$. The image of $C \in \mathcal{C}_{\mathbf{d}}$ under g is obtained by replacing all stubs with their corresponding nodes and consolidating the result as a multiset. The uniform distribution on $\mathcal{C}_{\mathbf{d}}$ induces a distribution on $\mathcal{G}_{\mathbf{d}}$ via g. Denote by $g^{-1} : \mathcal{G}_{\mathbf{d}} \to 2^{\mathcal{C}_{\mathbf{d}}}$ the function that assigns to each element of $\mathcal{G}_{\mathbf{d}}$ its preimage in $\mathcal{C}_{\mathbf{d}}$ under g.

Definition 21 (Configuration Model)**.** *Let $\lambda_{\mathbf{d}}$ be the uniform distribution on $\mathcal{C}_{\mathbf{d}}$. The* configuration model *on $\mathcal{G}_{\mathbf{d}}$ is the distribution $\mu_{\mathbf{d}} = \lambda_{\mathbf{d}} \circ g^{-1}$.*

The distinction between $\eta_{\mathbf{d}}$ and $\mu_{\mathbf{d}}$ – and its implications for data analysis – was recently highlighted by Fosdick et al. [76]. We have diverged from the terminology of the authors: our "uniform model" is their "configuration model on non-loopy, vertex-labeled multigraphs" and our "configuration model" is their "configuration model on non-loopy, stub-labeled multigraphs."

The distinction between uniform and configuration models lies in how they weight graphs with parallel edges. Let C_1 and C_2 be two configurations. Suppose that C_1 contains the matchings $(i_1, j_1), (i_2, j_2)$ and C_2 contains the matchings $(i_1, j_2), (i_2, j_1)$, and that they otherwise agree on all other stubs. Let $G = g(C_1) = g(C_2)$. Under the uniform model, G is considered to be a single state, weighted equally with all other states. Under the configuration model, on the other hand, the probability mass

placed on G is proportional to $|g^{-1}(G)|$, reflecting both C_1 and C_2 as distinct states. In particular, the configuration model $\mu_{\mathbf{d}}$ will tend to place higher probabilistic weight on elements of $\mathcal{G}_{\mathbf{d}}$ with large numbers of parallel edges than will the uniform model $\eta_{\mathbf{d}}$.

In the absence of parallel edges, the uniform and configuration models are closely related. Let A be the event that G is simple, without self-loops or parallel edges. Then, it is direct to show (e.g. [37]) that, for all G, $\eta_{\mathbf{d}}(G|A) = \mu_{\mathbf{d}}(G|A)$. The reason is that, when G is simple, the sizes of the preimages $g^{-1}(G)$ depend only on the degree sequence $\mathbf{d}$. Since $\mathbf{d}$ is fixed in $\mathcal{G}_{\mathbf{d}}$, these preimages all have the same size. Thus, when a *simple* random graph is required, the uniform model $\eta_{\mathbf{d}}$ and configuration model $\mu_{\mathbf{d}}$ are in principle interchangeable, in the sense that we can sample from $\eta_{\mathbf{d}}(\cdot|A)$ by repeatedly sampling from $\mu_{\mathbf{d}}$ until a simple graph is produced. Furthermore, when the degree sequence $\mathbf{d}$ grows slowly relative to n, $\mu_{\mathbf{d}}(A)$ is bounded away from zero by a function that depends on moments of $\mathbf{d}$ when n grows large [37, 137, 11]. This in turn provides an upper bound on the expected number of samples from $\mu_{\mathbf{d}}$ required to produce a single sample from $\eta_{\mathbf{d}}(\cdot|A)$. The computational importance of this relationship is that stub-matching for sampling from $\mu_{\mathbf{d}}$ is well-understood and often fast.

For dense graphs, $\mu_{\mathbf{d}}(A)$ may be extremely small, and the number of samples required to produce a simple graph may be prohibitive. While it is possible to make post-hoc edits to the graph to remove self-loops and multiple edges [138, 192], such methods can generate substantial and uncontrolled bias in finite graphs. Second and more importantly for our context, there is no equivalence between the unconditional distributions $\eta_{\mathbf{d}}$ and $\mu_{\mathbf{d}}$ on spaces of multigraphs. Stub-matching cannot therefore be used to sample from $\eta_{\mathbf{d}}$ when modeling considerations allow the presence of multiple edges.

5.2.1 Markov Chain Monte Carlo

An alternative approach to sampling uses Markov chains to explore structured sets of graphs. There exists a large constellation of related algorithms for this class of task,

including the sampling of marginal-constrained binary matrices [205, 13]; degree-regular [206, 135, 109] and degree-heterogeneous [45, 197, 32, 61] simple graphs; and graphs with degree-correlation constraints [6]. Most of these algorithms operate by repeatedly swapping edges in such a way as to preserve the required graph structure.

A fairly general variant, formulated by Fosdick et al. [76], can sample from either the uniform model $\eta_{\mathbf{d}}$ or the configuration model $\mu_{\mathbf{d}}$ on $\mathcal{G}_{\mathbf{d}}$. We define an edge swap to be a random function of two edges that share no nodes.[1] It interchanges a node on the first edge with a node on the second:

$$\text{EdgeSwap}((i,j),(k,\ell)) = \begin{cases} (i,k),(j,\ell) & \text{with probability } 1/2 \\ (i,\ell),(j,k) & \text{with probability } 1/2\,. \end{cases}$$

An edge swap does not change the total number of edges incident to nodes i, j, k, or ℓ, and therefore preserves $\mathbf{d}$. Starting from a graph $G_0 \in \mathcal{G}_{\mathbf{d}}$, repeated edge-swaps can therefore be used to obtain a random sequence of elements of $\mathcal{G}_{\mathbf{d}}$. Since each element of this sequence depends stochastically only on its predecessor, this sequence is a Markov chain. We perform Markov Chain Monte Carlo as follows. At each time step, we select two random edges (i,j) and (k,ℓ), uniformly selected from the set of pairs of edges with four distinct node indices. We then perform a pairwise edge-swap of these edges with *acceptance probability*

$$a((i,j),(k,\ell)) \triangleq \begin{cases} 1 & \text{configuration model } \mu_{\mathbf{d}} \\ (w_{ij}w_{k\ell})^{-1} & \text{uniform model } \eta_{\mathbf{d}}\,. \end{cases} \tag{5.3}$$

In the case that the edge-swap is not accepted, we record the current state again and

[1] Swaps involving edges that intersect are used when sampling from spaces that include self-loops [76].

resample. Formally,

Algorithm 5: MCMC Sampling for $\eta_{\mathbf{d}}$ and $\mu_{\mathbf{d}}$

Input: degree sequence $\mathbf{d}$, initial graph $G_0 \in \mathcal{G}_{\mathbf{d}}$, target distribution $\rho \in \{\eta_{\mathbf{d}}, \mu_{\mathbf{d}}\}$, sample interval $\delta t \in \mathbb{Z}_+$, sample size $s \in \mathbb{Z}_+$.

Initialization: $t \leftarrow 0$, $G \leftarrow G_0$

for $t = 1, 2, \ldots, s(\delta t)$ **do**

  sample (i, j) and (k, ℓ) uniformly at random from $\binom{E_t}{2}$

  if Uniform$([0, 1]) \leq a((i, j), (k, \ell))$ **then**

    $G_t \leftarrow \text{EdgeSwap}((i, j), (k, \ell))$

  else

    $G_t \leftarrow G_{t-1}$

Output: $\{G_t$ such that $t | \delta t\}$

For sufficiently large sample intervals δt, the output of Algorithm 5 will be approximately i.i.d. according to the target distribution ρ, as guaranteed by the following result.

Theorem 6 (Fosdick et al. [76])**.** *The Markov chain $\{G_t\}$ defined by Algorithm 5 is ergodic and reversible with respect to the input distribution ρ. As consequence, samples $\{G_t\}$ generated by Algorithm 5 are asymptotically independent and identically distributed according to ρ as $\delta t \to \infty$.*

These results provide a principled solution to the problem of asymptotically exact sampling from $\eta_{\mathbf{d}}$, and can therefore be used to construct an estimator $\hat{\mathbf{\Omega}}^{\text{mc}}$ of $\mathbf{\Omega}$, given by (5.1), with arbitrary levels of accuracy. It suffices to let the sample size s and sample interval δt grow large. There are two performance-related issues when using Algorithm 5 in practice, both of which are connected to the number of edges m. First is the question of how large δt should be to ensure that the samples are sufficiently close to independence. Heuristically, δt should scale with the mixing time of the chain, but very few bounds on mixing times for chains of this type appear to be available. In several recent papers, Greenhill [93, 92] and collaborators [70] have derived the only bounds known to this author for edge-swap Markov chains. In the space of simple graphs, under certain regularity conditions on the degree sequence, they

provide a mixing time bound with scaling $O({d_*}^{14} m^{10} \log m)$, where $d_* = \max_i d_i$. The scaling of this upper bound very poor, especially with regard to m, and is therefore not reassuring for practical applications. The second issue relates to the acceptance probabilities themselves. In a dense multigraph the number W_{ij} of edges between i and j will typically be large, resulting in low acceptance rates. Indeed, supposing that a typical entry W_{ij} scales approximately linearly with m, a typical acceptance probability would scale roughly as m^{-2}. A standard coupon-collector argument shows that it takes roughly $O(m \log m)$ accepted transitions to ensure that each edge has been swapped at least once, which would appear a reasonable requirement for a well-mixed chain. We therefore conjecture that the overall mixing time of Algorithm 5 for the uniform model on dense multigraphs is no smaller than $O(m^3 \log m)$, though a more precise statement and proof would be welcome. While much better than the best known proven results, such a scaling could likely be prohibitive for graphs of even modest size.

These considerations suggest that forming the MCMC estimate $\hat{\mathbf{\Omega}}^{\mathrm{mc}}$ may not be a computationally practical way to estimate $\mathbf{\Omega}$ when m is large. Despite these limitations, Algorithm 5 lies at the heart of our main results in the next section.

5.3 A Dynamical Approach to Model Moments

We introduce some additional notation to facilitate calculations. The transpose of vector $\mathbf{u}$ is denoted $\mathbf{u}^T$, and the inner product of $\mathbf{u}$ and $\mathbf{v}$ by $\mathbf{u}^T\mathbf{v}$. We denote the ith row or column of matrix $\mathbf{W}$ by $\mathbf{w}_i$; all matrices we encounter will be symmetric and so no ambiguity will arise. Let $\mathbf{e}$ be the vector of ones; the dimension of $\mathbf{e}$ will be clear in context. Similarly, let $\mathbf{e}_i$ be the ith standard basis vector. All sums over node indices i, j, k, ℓ have implicit limits from 1 to n. Finally, $a \wedge b$ and $a \vee b$ denote the pairwise minimum and maximum of scalars a and b, respectively.

Algorithm 5 describes a stochastic dynamical update on the space $\mathcal{G}_{\mathbf{d}}$ of multigraphs, which we identify with the space of symmetric matrices with nonnegative integer entries and zero diagonals. Let $\mathbf{\Delta}(t) = \mathbf{W}(t+1) - \mathbf{W}(t)$ be the (random)

increment in $\mathbf{W}$ in timestep $t+1$. We implicitly regard $\mathbf{W}$ and $\mathbf{\Delta}$ as functions of t, suppressing the argument for notational sanity when there is no possibility of confusion. We can separate $\mathbf{\Delta} = \mathbf{\Delta}^+ - \mathbf{\Delta}^-$, where $\Delta_{ij}^+ = (\Delta_{ij} \vee 0)$ and $\Delta_{ij}^- = (-\Delta_{ij} \vee 0)$. The first term Δ_{ij}^+ describes the (random) number of edges flowing into the pair (i,j) and the second term Δ_{ij}^- the random number of edges flowing out. Conservation of edges implies that $\sum_{ij} \Delta_{ij}^+ = \sum_{ij} \Delta_{ij}^-$. Since a pair of nodes can only gain or lose one edge at a time under the dynamics, the entries Δ_{ij}^+ and Δ_{ij}^- are Bernoulli random variables. These Bernoulli variables are not independent, since at most two entries of each matrix are nonzero in a given timestep. Let $\boldsymbol{\delta}^+ = \mathbb{E}[\mathbf{\Delta}^+]$ and $\boldsymbol{\delta}^- = \mathbb{E}[\mathbf{\Delta}^-]$.

Two things must hold at stationarity of Algorithm 5. First, all moments of $\mathbf{W}$ must be constant in time. Second, since the stationary distribution of Algorithm 5 is the target distribution ρ by construction, these moments of $\mathbf{W}$ are the desired moments of ρ. We can therefore approximately compute moments of ρ by approximately solving conveniently chosen stationarity conditions. A useful set is given by

$$\mathbb{E}[w_{ij}(t+1)^p - w_{ij}(t)^p] = \mathbb{E}\left[(w_{ij}(t) + \Delta_{ij}(t))^p - w_{ij}(t)^p\right] = 0\,, \tag{5.4}$$

for positive integers p. These equations express directly the time-invariance of the moments $\mathbb{E}[w_{ij}(t)^p]$ at stationarity.

5.3.1 Illustration: The Configuration Model

We will derive a version of the Chung-Lu estimator $\hat{\mathbf{\Omega}}^0$ for the configuration model by studying (5.4) when $p = 1$.

Theorem 7. *Under the configuration model $\mu_{\mathbf{d}}$, for all $i \neq j$,*

$$\omega_{ij} = \frac{d_i d_j - \mathbb{E}[\mathbf{w}_i^T \mathbf{w}_j] - \mathbb{E}[w_{ij}^2]}{2m - d_i - d_j} \tag{5.5}$$

Proof. We first derive expressions for $\mathbf{\Delta}^-$ and $\mathbf{\Delta}^+$ by stepping through the stages of Algorithm 5. For the former, note that $\Delta_{ij}^- = 1$ only if edge (i,j) is sampled in the first stage of the iteration. The probability that edges (i,j) and (k,ℓ) are sampled,

assuming that all four indices are distinct, is $z(\mathbf{W})^{-1}W_{ij}W_{k\ell}$, where

$$z(\mathbf{W}) = \sum_{\substack{i,j \\ k,\ell \notin \{i,j\}}} W_{ij}W_{k\ell}$$

gives the total number of ways to pick two edges with four distinct indices. Under the configuration model, $a((i,j),(k,\ell)) = 1$. Summing across k and ℓ and taking expectations, we obtain

$$\begin{aligned}\delta_{ij}^{-} &= \frac{1}{z(\mathbf{W})}\mathbb{E}\left[\sum_{k,\ell \notin \{i,j\}} w_{ij}w_{k\ell}\right] \\ &= \frac{1}{z(\mathbf{W})}\mathbb{E}\left[w_{ij}\left(\sum_{k,\ell} w_{k\ell} - \sum_{k}(w_{ki}+w_{kj}) - \sum_{\ell}(w_{i\ell}+w_{j\ell}) + 3w_{ij}\right)\right].\end{aligned}$$

Recalling constraints such as $\sum_{k,\ell} W_{k\ell} = 2m$ and $\sum_k W_{ki} = d_i$, this expression simplifies to

$$\delta_{ij}^{-} = \frac{1}{z(\mathbf{W})}\left(2\omega_{ij}(m - d_i - d_j) + 3\mathbb{E}[w_{ij}^2]\right)$$

We can derive a similar expression for δ_{ij}^{+}. Fix two additional indices k and ℓ, such that all four indices i,j,k,ℓ are distinct. A new edge (i,j) can be generated from selecting for swap either of the pairs $\{(i,k),(\ell,j)\}$ or $\{(i,\ell),(k,j)\}$. These events occur with probabilities $z(\mathbf{W})^{-1}w_{ik}w_{\ell j}$ and $z(\mathbf{W})^{-1}w_{i\ell}w_{kj}$, respectively. Having selected edges $\{(i,k),(\ell,j)\}$, edges $\{(i,j),(k,\ell)\}$ are formed by the swap with probability $\frac{1}{2}$; otherwise $\{(i,\ell),(k,j)\}$ are formed. Summing across k and ℓ and computing expectations,

we have

$$\begin{aligned}
\delta_{ij}^{+} &= \frac{1}{2z(\mathbf{W})}\mathbb{E}\left[\sum_{\substack{k,\ell\notin\{i,j\}\\ k\neq\ell}} w_{ik}w_{\ell j} + \sum_{\substack{k,\ell\notin\{i,j\}\\ k\neq\ell}} w_{i\ell}w_{kj}\right] \\
&= \frac{1}{z(\mathbf{W})}\mathbb{E}\left[\sum_{\substack{k,\ell\notin\{i,j\}\\ k\neq\ell}} w_{ik}w_{\ell j}\right] \\
&= \frac{1}{z(\mathbf{W})}\mathbb{E}\left[\left(\sum_{k} w_{ik}\right)\left(\sum_{\ell} w_{\ell j}\right) - w_{ij}\sum_{k}\left(w_{ik}+w_{jk}\right) - \sum_{k} w_{ik}w_{kj} + w_{ij}^2\right] \\
&= \frac{1}{z(\mathbf{W})}\left(d_i d_j \ - \omega_{ij}(d_i+d_j) - \mathbb{E}[\mathbf{w}_i^T\mathbf{w}_j] + \mathbb{E}[w_{ij}^2]\right) . \qquad (5.6)
\end{aligned}$$

Choosing $p = 1$ in (5.4), we must have $\delta_{ij}^{+} = \delta_{ij}^{-}$ at stationarity. Inserting our derived expressions and solving for ω_{ij} yields the result. □

Theorem 7 does not give an explicit operational solution for ω_{ij}, since the right-hand side contains higher moments of $\mathbf{W}$. Progress can be made in the "large, sparse regime," in which we assume that n is large and the entries of $\mathbf{W}$ and $\mathbf{d}$ small relative to m. Recalling that $\hat{\omega}_{ij}^0 = \frac{d_i d_j}{2m}$, we can rewrite (5.5) as

$$\omega_{ij} = \left(1 - \frac{d_i + d_j}{2m}\right)^{-1}\left(1 - \frac{\mathbb{E}[\mathbf{w}_i^T\mathbf{w}_j] - \mathbb{E}[w_{ij}^2]}{d_i d_j}\right)\hat{\omega}_{ij}^0.$$

In the large, sparse heuristic, each entry of $\mathbf{d}$ is small in comparison to m, and the first error factor is near unity. Similarly, the expression $\mathbb{E}[\mathbf{w}_i^T\mathbf{w}_j] - \mathbb{E}[w_{ij}^2]$ implicitly contains up to $n-1$ nonzero products of entries of $\mathbf{W}$. On the other hand, the denominator contains $(n-1)^2$ such terms, and we therefore expect the second error factor to also lie near unity. We therefore expect that $\omega_{ij} \to \hat{\omega}_{ij}^0$ "in the large, sparse regime.". This statement can be made precise by specifying the asymptotic behavior of $\mathbf{W}$ with respect to n, which is beyond our present scope. Through analysis of Algorithm 5, we have derived both $\hat{\mathbf{\Omega}}^0$ and explicit error terms that are often elided in the network science literature.

5.3.2 Moments of the Uniform Model

The analysis of the uniform model is somewhat more subtle. As seen in the configuration model above, many of the sums that appear in the calculations of $\boldsymbol{\delta}^+$ and $\boldsymbol{\delta}^-$ reduced to fixed constants due to the degree constraints. Unfortunately, there is no analogous simplification in the uniform model $\eta_{\mathbf{d}}$. Because of this, we require some additional technology in order to make progress.

Define the binary matrix $\mathbf{X} \in \{0,1\}^{n \times n}$ entrywise by $x_{ij} = \mathbb{1}(w_{ij} = 0)$. For convenience, we adopt the convention $0/0 = 0$ under which the identity $w_{ij}/w_{ij} = x_{ij}$ holds even when $w_{ij} = 0$. We can interpret $\mathbf{X}$ as the adjacency matrix of the simple graph obtained by collapsing all sets of parallel edges in a multigraph into single edges. Let $\mathbf{b} = \mathbf{X}\mathbf{e}$, the vector of row sums of $\mathbf{X}$. The vector $\mathbf{b}$ is interpretable as the collapsed degree sequence, whose ith entry gives the number of distinct neighbors of node i. Let $y = \frac{1}{2}\mathbf{e}^T\mathbf{b}$ give the total number of collapsed edges. The expectations of $\mathbf{X}$, $\mathbf{b}$, and y play important roles in our analysis. We denote them

$$\boldsymbol{\chi} = \mathbb{E}[\mathbf{X}] \,, \quad \boldsymbol{\beta} = \mathbb{E}[\mathbf{b}] \,, \text{ and } \quad \psi = \mathbb{E}[y] \,.$$

The objects $\boldsymbol{\chi}$, $\boldsymbol{\beta}$, and ψ are all implicitly deterministic functions of $\mathbf{d}$. Throughout this section, we let $I = (i_1, \ldots, i_p)$ be a set of p not-necessarily-distinct indices, and let $K = ((k_1, \ell_1), \ldots, (k_q, \ell_q))$ be a set of q not-necessarily-distinct dyadic indices. If $\mathbf{v}$ is a vector, we let $\mathbf{v}_I$ denote the vector with entries $(v_{i_1}, \ldots, v_{i_p})$. Similarly, if $\mathbf{A}$ is a matrix, we let $\mathbf{A}_K$ denote the vector with entries $(a_{k_1,\ell_1}, \ldots, a_{k_q,\ell_q})$.

We first require control over the behavior of $\boldsymbol{\beta}$ with respect to $\mathbf{d}$.

Definition 22 (Regularity Constant for $\boldsymbol{\beta}$)**.** *Let $u(\mathbf{d})$ be the smallest real number such that, for any degree sequence $\mathbf{d}'$ such that $\mathbf{d}' \geq \mathbf{d}$ entrywise, and for all distinct indices i and j,*

$$\|\boldsymbol{\beta}(\mathbf{d}' + \mathbf{e}_i + \mathbf{e}_j) - \boldsymbol{\beta}(\mathbf{d}')\|_\infty \leq u(\mathbf{d}) \,.$$

Intuitively, $u(\mathbf{d})$ provides a bound on the sensitivity of $\boldsymbol{\beta}$ to increments in entries

of the degree sequence. Indeed, $u(\mathbf{d})/\sqrt{2}$ is by definition the Lipschitz constant (with respect to the ℓ^2 and ℓ^∞ norms) for the restriction of $\boldsymbol{\beta}$ to the set $\{\mathbf{d}' : \mathbf{d}' \geq \mathbf{d}\}$. Since $0 \leq \beta_\ell(\mathbf{d}) \leq n-1$, we have trivially that $u(\mathbf{d}) \leq n-1$. To produce a lower bound, we can also produce degree sequences $\mathbf{d}$ such that $u(\mathbf{d}) \geq 1$. To see this, note that, if $d_i = d_j = 0$, then $\beta_i(\mathbf{d}) = \beta_j(\mathbf{d}) = 0$. On the other hand, $\beta_i(\mathbf{d}+\mathbf{e}_i+\mathbf{e}_j) = \beta_j(\mathbf{d}+\mathbf{e}_i+\mathbf{e}_j) = 1$.

We also define v as the smallest real number such that, for all i and j,

$$|\psi(\mathbf{d}+\mathbf{e}_i+\mathbf{e}_j) - \psi(\mathbf{d})| \leq v(\mathbf{d}) \,.$$

Similarly to the above, we have the trivial bounds $1 \leq v(\mathbf{d}) \leq n(n-1)$, since $n(n-1)$ is the largest possible number of nonzero entries of $\mathbf{X}$.

Conjecture 1. *For all* $\mathbf{d}$*, we have* $u(\mathbf{d}) \leq 1$ *and* $v(\mathbf{d}) \leq 1$*.*

The intuition behind this conjecture is as follows. If $G \in \mathcal{G}_{\mathbf{d}}$, the sequence $\mathbf{d}+\mathbf{e}_i+\mathbf{e}_j$ can be instantiated by a graph $G' \in \mathcal{G}_{\mathbf{d}+\mathbf{e}_i+\mathbf{e}_j}$ in which a single edge has been added between nodes i and j. Trivially, this operation does not decrease the degrees of any nodes, does not increase the degrees of any nodes by more than one, and does not increase the total number of edges by more than one. Conjectures 1 states that the same is true of the expected collapsed degrees $\boldsymbol{\beta}$ and expected number of collapsed edges ψ. The bounds we present below do not formally depend on the truth of Conjectures 1, but they are not guaranteed to be meaningful unless u and v are indeed small in comparison to n. Unfortunately, the complex combinatorial structure of $\mathcal{G}_{\mathbf{d}}$ renders a proof of our conjecture obscure, and we leave such a proof to future work. In Section 5.5.1, we will show anecdotal numerical experiments consistent with this conjecture. For now, we define

$$u_*(\mathbf{d}) = \max\{u(\mathbf{d}), 1\} \quad \text{and} \quad v_*(\mathbf{d}) = \max\{v(\mathbf{d}), 1\} \,.$$

Conjectures 1 then states that $u_*(\mathbf{d}) = v_*(\mathbf{d}) = 1$ for all $\mathbf{d}$.

Theorem 8. *Suppose that* $\eta_{\mathbf{d}}(\mathbf{X}_K = \mathbf{e}) > 0$. *Then, for any* $i \in [n]$,

$$|\mathbb{E}[b_i|\mathbf{X}_K = \mathbf{e}] - \beta_i(\mathbf{d})| \le 2qu_*(\mathbf{d}) \ . \tag{5.7}$$

Additionally,

$$|\mathbb{E}[y|\mathbf{X}_K = \mathbf{e}] - \psi(\mathbf{d})| \le 2qv_*(\mathbf{d}) \ . \tag{5.8}$$

Proof. We will prove (5.7); the proof of (5.8) is parallel. Fix $k, \ell \in [n]$. The distribution $\eta_{\mathbf{d}+\mathbf{e}_k+\mathbf{e}_\ell}$ is supported on graphs with n nodes and $m+1$ edges. Let us condition $\eta_{\mathbf{d}+\mathbf{e}_k+\mathbf{e}_\ell}$ on the event $w_{k\ell} \ge 1$. Then, there exists at least one edge (k, ℓ). Since $\eta_{\mathbf{d}+\mathbf{e}_k+\ell}$ is itself uniform, the conditioned distribution, in which the remaining m edges, is also uniform. Indeed, since we have already assigned an edge incident to nodes k and ℓ, the conditional distribution is uniform over configurations of the remaining m edges in which the degrees sum to $\mathbf{d}$. This is exactly $\eta_{\mathbf{d}}$, and we therefore obtain the identity

$$\eta_{\mathbf{d}}(G) = \eta_{\mathbf{d}+\mathbf{e}_k+\mathbf{e}_\ell}(G \uplus \{(k, \ell)\} | w_{k\ell} \ge 1) \ , \tag{5.9}$$

where $\uplus$ denotes multiset union. This identity relates the operations of conditioning and degree sequence modification. Now, let $\mathbf{v}(K) = \sum_{s=1}^{q}(\mathbf{e}_{k_s} + \mathbf{e}_{\ell_s})$. By iterating (5.9), we obtain

$$\eta_{\mathbf{d}}(G) = \eta_{\mathbf{d}+\mathbf{v}(K)}\left(G \uplus \biguplus_{s=1}^{q}\{k_s, \ell_s\} \middle| \mathbf{W}_K \ge \mathbf{e}\right) \ . \tag{5.10}$$

Note that the conditioning event can be equivalently written $\mathbf{X}_K = \mathbf{e}$.

For the remainder of this proof, let the symbol $\mathbb{E}_{\mathbf{z}}$ denote expectations with respect to $\eta_{\mathbf{z}}$. We use (5.9) to estimate $\mathbb{E}_{\mathbf{d}+\mathbf{v}(K)}[b_i|\mathbf{X}_K = \mathbf{e}]$ via a two-step experiment. We first sample $G \sim \eta_{\mathbf{d}}$ and compute b_i. We then add the q edges $\{(k_s, \ell_s)\}$ sequentially. Doing so does not decrease b_i, and can increase b_i by no more than $q \le qu_*$. Taking

expectations, we obtain the bound

$$\beta_i(\mathbf{d}) \le \mathbb{E}_{\mathbf{d}+\mathbf{v}(K)}[b_i|\mathbf{X}_K = \mathbf{e}] \le \beta_i(\mathbf{d}) + qu_* \ .$$

Applying (1) inductively, we also have

$$\beta_i(\mathbf{d}) - qu_* \le \beta_i(\mathbf{d} + \mathbf{v}(K)) \le \beta_i(\mathbf{d}) + qu_* \ .$$

We infer that

$$\left|\mathbb{E}_{\mathbf{d}+\mathbf{v}(K)}[b_i|\mathbf{X}_K = \mathbf{e}] - \beta_i(\mathbf{d} + \mathbf{v}(K))\right| \le 2qu_* \ .$$

Since $\mathbf{d}$ and K were arbitrary, we can absorb $\mathbf{v}(K)$ into $\mathbf{d}$, obtaining the required statement

$$|\mathbb{E}_{\mathbf{d}}[b_i|\mathbf{X}_K = \mathbf{e}] - \beta_i(\mathbf{d})| \le 2qu_* \ ,$$

The only subtlety in this case is that the expectation must exist. For this it is sufficient that $\eta_{\mathbf{d}}(\mathbf{X}_K = \mathbf{e}) > 0$, as assumed by hypothesis. □

Theorem 8 is our primary tool for proving second-moment bounds on the entries of $\mathbf{b}$. From this point forward, we will assume that $\mathbf{d}$ is fixed. The symbols η, $\mathbb{E}$ will refer to the uniform model with degree sequence $\mathbf{d}$ and expectations with respect to that model, respectively.

Lemma 3. *Let $i \ne j \ne k$. The following bounds hold.*

$$|\mathbb{E}[b_i x_{jk}] - \beta_i \chi_{jk}| \le 2u_* \chi_{jk} \tag{5.11}$$

$$|\mathbb{E}[y x_{jk}] - \psi \chi_{jk}| \le 2v_* \chi_{jk} \tag{5.12}$$

$$|\mathbb{E}[b_i b_j] - \beta_i \beta_j| \le 2u_* (\beta_i \wedge \beta_j) \tag{5.13}$$

$$\mathbb{E}[b_i^2 b_j^2] \le (\beta_i + 6u_*)^2 (\beta_j + 6u_*)^2 \tag{5.14}$$

$$\mathrm{var}\,(y) \le 2v_* \psi \ . \tag{5.15}$$

Proof. To prove (5.11), write

$$\mathbb{E}[b_i x_{jk}] = \eta(x_{jk} = 1)\mathbb{E}[b_i | x_{jk} = 1] = \chi_{jk}\mathbb{E}[b_i | x_{jk} = 1]$$

and apply Theorem 8. To prove (5.12), we similarly write

$$\mathbb{E}[y x_{jk}] = \eta(x_{jk} = 1)\mathbb{E}[y | x_{jk} = 1] = \chi_{jk}\mathbb{E}[y | x_{jk} = 1]$$

and apply Theorem 8. To prove (5.13), write

$$\mathbb{E}[b_i b_j] = \sum_{\ell} \mathbb{E}[b_i x_{j\ell}]$$

Now applying (5.11), we obtain

$$|\mathbb{E}[b_i b_j] - \beta_i \beta_j| \leq 2u_* \sum_{\ell} \chi_{j\ell} = 2u_* \beta_j \ .$$

Since we could have expanded b_j instead of b_i, we can choose the smaller of these, and the result follows. The proof of (5.14) is similar. We expand the sums and apply Theorem 8. The first step is

$$\begin{aligned}
\mathbb{E}[b_i^2 b_j^2] &= \sum_{k,\ell,h} \mathbb{E}[x_{ik} x_{i\ell} x_{jh}]\mathbb{E}[b_j | x_{ik} x_{i\ell} x_{jh} = 1] \\
&\leq \sum_{k,\ell,h} \mathbb{E}[x_{ik} x_{i\ell} x_{jh}](\beta_j + 6u_*) \\
&= (\beta_j + 6u_*)\mathbb{E}[b_i^2 b_j] \ .
\end{aligned}$$

Repeating this procedure three more times proves the result. Finally, to prove (5.15), write

$$\text{var}\,(y) = \mathbb{E}[y^2] - \psi^2 = \frac{1}{2}\sum_{ij} \chi_{ij}\mathbb{E}[y | x_{ij} = 1] - \psi^2 \leq \frac{1}{2}\sum_{ij} \chi_{ij}(\psi + 2v_*) - \psi^2 = 2v_*\psi \ .$$

We have used (5.12) in the inequality. □

Lemma 4. *We have*

$$\left|\delta_{ij}^- - \frac{2\psi\chi_{ij}}{z(\mathbf{W})}\right| \le \frac{\epsilon_{ij}^-}{z(\mathbf{W})} \quad \textit{and} \quad \left|\delta_{ij}^+ - \frac{\beta_i\beta_j}{z(\mathbf{W})}\right| \le \frac{\epsilon_{ij}^+}{z(\mathbf{W})} \,,$$

where

$$\begin{aligned}
\epsilon_{ij}^- &\triangleq \chi_{ij}(\beta_i + \beta_j + 3 + 4v_* + 2u_*) \\
\epsilon_{ij}^+ &\triangleq \chi_{ij}(\beta_i + \beta_j + 4u_* - 1) + (2u_* + 1)(\beta_i \wedge \beta_j) \,.
\end{aligned}$$

Proof. We require expressions for $\mathbf{\Delta}^-$ and $\mathbf{\Delta}^+$. These are as in the calculation for the configuration model in Section 5.3.1, except that there now appears an acceptance probability $a((i,j),(k,\ell)) = \frac{1}{w_{ij},w_{k\ell}}$ that modifies the swap probabilities. The acceptance probability has the effect of replacing instances of w_{ij} with x_{ij}. Performing the algebra and simplifying, we find that

$$\delta_{ij}^- = \frac{1}{z(\mathbf{W})}\mathbb{E}\left[2x_{ij}(y - b_i - b_j) + 3x_{ij}\right] \tag{5.16}$$

$$\delta_{ij}^+ = \frac{1}{z(\mathbf{W})}\mathbb{E}\left[b_ib_j - x_{ij}(b_i + b_j) - \mathbf{x}_i^T\mathbf{x}_j + x_{ij}\right] \,. \tag{5.17}$$

These are indeed the same expressions as in the configuration model in Section 5.3.1, with $\mathbf{W}$ replaced by $\mathbf{X}$. We have used the identity $x_{ij}^2 = x_{ij}$. Computing the expectation of the first line yields

$$\delta_{ij}^- = \frac{1}{z(\mathbf{W})}\left(2\mathbb{E}[yx_{ij}] - 2\mathbb{E}[b_ix_{ij}] - 2\mathbb{E}[b_jx_{ij}] + 3\chi_{ij}\right) \,.$$

Applying (5.11) and (5.12), we obtain the bound

$$\left|\delta_{ij}^- - \frac{1}{z(\mathbf{W})}\left(\chi_{ij}(2(\psi - \beta_i - \beta_j) + 3)\right)\right| \le \frac{4}{z(\mathbf{W})}\chi_{ij}(v_* + 2u_*) \,.$$

We similarly compute

$$\delta_{ij}^+ = \frac{1}{z(\mathbf{W})}\left(\mathbb{E}[b_ib_j] - \mathbb{E}[b_ix_{ij}] - \mathbb{E}[b_jx_{ij}] - \mathbb{E}[\mathbf{x}_i^T\mathbf{x}_j] + \chi_{ij}\right) \,.$$

We note that, since $\mathbf{X}$ is binary, $0 \le \mathbf{x}_i^T \mathbf{x}_j \le b_i \wedge b_j$, and therefore $0 \le \mathbb{E}[\mathbf{x}_i^T \mathbf{x}_j] \le \beta_i \wedge \beta_j$. Applying this observation in concert with (5.13) and (5.11), we find

$$\left| \delta_{ij}^+ - \frac{1}{z(\mathbf{W})} \left(\beta_i \beta_j - \chi_{ij} \beta_i - \chi_{ij} \beta_j + \chi_{ij} \right) \right| \le \frac{1}{z(\mathbf{W})} \left((2u_* + 1)(\beta_i \wedge \beta_j) + 4u_* \chi_{ij} \right) .$$

Moving the unwanted terms to the righthand side in both bounds proves the lemma. □

Theorem 9 (Expectations of $\mathbf{X}$)**.** *We have*

$$\left| \chi_{ij} - \frac{\beta_i \beta_j}{2\psi} \right| \le \epsilon_{ij}(\boldsymbol{\beta}) \triangleq \frac{\epsilon_{ij}^+(\boldsymbol{\beta}) + \epsilon_{ij}^-(\boldsymbol{\beta})}{2\psi} .$$

Furthermore,

$$\epsilon_{ij}(\boldsymbol{\beta}) = \frac{2\chi_{ij}(\beta_i + \beta_j + 3u_* + 2v_* + 2) + (2u_* + 1)(\beta_i \wedge \beta_j)}{2\psi} .$$

Proof. Setting $p = 1$ in (5.4) again yields $\delta_{ij}^+ - \delta_{ij}^- = 0$. Applying Lemma 4 and the triangle inequality, we obtain

$$\left| \frac{\beta_i \beta_j}{z(\mathbf{W})} - \frac{2\psi \chi_{ij}}{z(\mathbf{W})} \right| \le \frac{\epsilon_{ij}^- + \epsilon_{ij}^+}{z(\mathbf{W})}$$

Multiplying through by $\frac{z(\mathbf{W})}{2\psi}$ proves the first claim. The expression for ϵ_{ij} is obtained inserting the expressions for ϵ_{ij}^- and ϵ_{ij}^+ from Lemma 4 and simplifying. □

Theorem 9 provides an asymptotic error bound of the form

$$\chi_{ij} = \frac{\beta_i \beta_j}{2\psi} \left(1 + O\left(\frac{\chi_{ij} v_*}{\beta_i \beta_j} + \chi_{ij} \frac{\beta_i + \beta_j}{\beta_i \beta_j} + \frac{u_*}{\beta_i \vee \beta_j} \right) \right) \tag{5.18}$$

as β_i and β_j grow large. This bound is admittedly relatively loose, even assuming that u_* and v_* are indeed small. In light of the numerical results presented below, we conjecture that much better bounds may be possible. This appears to be a promising direction for future work.

We can recognize the leading term in (5.18):

$$f_{ij}(\boldsymbol{\beta}) = \frac{\beta_i \beta_j}{2\psi} ,$$

the same functional form f_{ij} as in the CL estimator defined in (5.2). Speaking somewhat figuratively, we can interpret Theorem 9 as indicating that $\mathbf{X}$, the matrix of the projected simple graph, approximately agrees in expectation with the Chung-Lu model (on off-diagonal entries) with parameter vector $\boldsymbol{\beta}$. However, tt would be incorrect to state that $\mathbf{X}$ is distributed according to any model that deterministically preserves a collapsed degree sequence. First, $\boldsymbol{\beta}$ does not in general possess integer entries. Second the collapsed degrees b_i are still stochastic, preserved only approximately in expectation.

5.3.3 First Moments of W

In the case of the configuration model, approximately solving the $p = 1$ stationarity condition yielded an approximation for $\boldsymbol{\Omega}$ in terms of the known vector $\mathbf{d}$. However, in the uniform model we derived an approximation only for $\boldsymbol{\chi}$ in terms of the unknown vector $\boldsymbol{\beta}$. Computing another equilibrium condition will allow us to both estimate $\boldsymbol{\Omega}$ from $\boldsymbol{\chi}$ and estimate $\boldsymbol{\beta}$ from $\mathbf{d}$. Take $p = 2$ in (5.4), obtaining

$$2\mathbb{E}[w_{ij}\Delta_{ij}] + \mathbb{E}[\Delta_{ij}^2] = 0 . \tag{5.19}$$

Study of this condition yields the following result.

Theorem 10. *Assume that* $f_{ij}(\boldsymbol{\beta}) < 1$. *Then,*

$$\left| \omega_{ij} - \frac{f_{ij}(\boldsymbol{\beta})}{1 - f_{ij}(\boldsymbol{\beta})} \right| \leq \frac{1}{1 - f_{ij}(\boldsymbol{\beta})} \left(\frac{2\epsilon'_{ij}(\boldsymbol{\beta}) + \epsilon_{ij}(\boldsymbol{\beta})}{2\psi} + \frac{\epsilon_{ij}(\boldsymbol{\beta})}{2} \right) ,$$

where $\epsilon_{ij}(\boldsymbol{\beta})$ *is as in Theorem 9 and*

$$\epsilon'_{ij}(\boldsymbol{\beta}) \triangleq \frac{2u_*}{\beta_i \vee \beta_j} + \frac{\sigma_{ij}}{\omega_{ij}} \frac{\sqrt{(\beta_i + 6u_*)^2(\beta_j + 6u_*)^2 - (\beta_i\beta_j - 2u_*(\beta_i \wedge \beta_j))^2}}{\beta_i\beta_j} + \sigma_{ij}\sqrt{2v_*\psi_i} + \omega_{ij}(\beta_i + \beta_j) + \sigma_{ij}\sqrt{2u_*}(\sqrt{\beta_i} + \sqrt{\beta_j})] \, .$$

The proof of Theorem 10 proceeds similarly to that of Theorem 9, albeit with more involved algebra. It is provided in the Supplementary Information. We note that, while it is notationally convenient to leave the final (inside the square root) term unexpanded, the term $\beta_i^2\beta_j^2$ cancels. The entire expression is therefore of polynomial order $-\frac{1}{2}$ in the entries of $\boldsymbol{\beta}$, and again goes to zero as these entries grow large.

Informally, Theorem 10 states that

$$\omega_{ij} \approx \frac{f_{ij}(\boldsymbol{\beta})}{1 - f_{ij}(\boldsymbol{\beta})} \, . \tag{5.20}$$

Recall that $f_{ij}(\boldsymbol{\beta}) \approx \chi_{ij}$ by Theorem 9, and that $\chi_{ij} = \eta(w_{ij} \geq 1)$ by definition. Then, (10) states that ω_{ij} is approximately equal to the odds that there is at least one edge present between nodes i and j. As we will see, this approximation gives us a method to compute the vector $\boldsymbol{\beta}$ in terms of the vector $\mathbf{d}$, thereby obtaining an approximation for the moments of $\mathbf{W}$. As in Theorem 9, the derived bounds are relatively loose, and substantially better ones may perhaps be obtained from further analysis.

5.3.4 Second Moments

Before proceeding, we briefly comment on the $p = 3$ stationarity condition. From this case on, it becomes quite tedious to control the error terms associated with factoring expectations. Omitting them, we obtain the approximation

$$\mathbb{E}[w_{ij}^2] \approx \omega_{ij}\left(\omega_{ij} + \frac{1}{1 - \chi_{ij}}\right) \, .$$

It follows that

$$\sigma_{ij}^2 = \operatorname{var}(w_{ij}) \approx \frac{\chi_{ij}}{(1-\chi_{ij})^2} \approx \omega_{ij}(\omega_{ij}+1)\,. \tag{5.21}$$

Note that, under this approximation, $\sigma_{ij}^2 > \omega_{ij}$ whenever $\chi_{ij} > 0$. It is common to model the entries of the adjacency matrix as Poisson random variables, for which the mean and variance are equal. The formula (5.21) suggests that this approach will be approximately correct for the uniform model when $\omega_{ij} \ll 1$, but systematically underestimate the variance for larger values.

5.4 Estimation of $\boldsymbol{\beta}$

We now possess approximate formulae for the low-order moments of $\mathbf{W}$ in terms of the vector $\boldsymbol{\beta}$. In practice, we do not observe $\boldsymbol{\beta}$ and must therefore estimate it from $\mathbf{d}$. To do so, we impose the degree constraint $\sum_j \omega_{ij} = d_i$ and insert the approximation given by Theorem 10. Eliding the error terms, we obtain

$$d_i \approx \sum_j \frac{f_{ij}(\boldsymbol{\beta})}{1-f_{ij}(\boldsymbol{\beta})}\,.$$

We therefore define the function $\mathbf{h} : \mathbb{R}_+^n \to \mathbb{R}^n$ componentwise as

$$h_i(\boldsymbol{\beta}) \triangleq \sum_j \frac{f_{ij}(\boldsymbol{\beta})}{1-f_{ij}(\boldsymbol{\beta})} \tag{5.22}$$

and aim to solve the equation

$$\mathbf{h}(\boldsymbol{\beta}) = \mathbf{d} \tag{5.23}$$

for $\boldsymbol{\beta}$. We define the estimator $\hat{\boldsymbol{\beta}}$ as the solution of (5.23). We then use the estimators $\hat{\chi}_{ij} \triangleq f_{ij}(\hat{\boldsymbol{\beta}})$ and $\hat{\omega}_{ij}^1 \triangleq \frac{f_{ij}(\hat{\boldsymbol{\beta}})}{1-f_{ij}(\hat{\boldsymbol{\beta}})}$ supplied by Theorems 9 and 10 as estimates of the moments of $\mathbf{W}$. In general, $\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$, since we have discarded the error terms derived in the previous section. We should therefore expect that $\hat{\boldsymbol{\beta}}$ is a biased estimator of $\boldsymbol{\beta}$,

and that $\hat{\mathbf{\Omega}}^{1}$ is a biased estimator of $\mathbf{\Omega}$. Experiments, however, will show that these biases are substantially smaller than those of $\hat{\mathbf{\Omega}}^{0}$.

To get some intuition on the behavior of (5.23), it is useful to consider two contrasting cases. First, consider the degree sequence $\mathbf{d} = d\mathbf{e}$. In this case, $\mathcal{G}_{\mathbf{d}}$ is the set of regular graphs in which all nodes have the same degree d. We can find a solution of (5.23) analytically. We assume that $\boldsymbol{\beta} = \beta\mathbf{e}$ for some scalar β. Then, (5.23) reads

$$\frac{(n-1)\beta^2}{n\beta - \beta^2} = d\,.$$

Solving for β yields the estimator $\hat{\beta}$:

$$\hat{\beta} = \frac{d}{1 + n^{-1}(d-1)}\,.$$

We see that, in a sparse limit in which we let $n \to \infty$ while $d = o(n)$, $\hat{\beta} \to d$. This reflects the asymptotic equivalence of uniform and configuration models under large, sparse limits.

Our second example illustrates a case in which no interpretable solution to (5.23) exists. Consider the star graph, which possess $k \geq 2$ leaves (labeled 1 through k) and a central node (labeled $k+1$). A single edge connects each leaf to node $k+1$. Node $k+1$ has degree k, while each leaf has degree 1. There are no valid edge-swaps, and the corresponding null space $\mathcal{G}_{\mathbf{d}}$ therefore contains only one element. We can thus read off the correct expected collapsed degree sequence: $\beta_{k+1} = k$ and $\beta_j = 1$ for $1 \leq j \leq k$. However, this sequence does not solve (5.23). Indeed, letting β_L denote the unknown shared collapsed degree for each leaf and β_C the collapsed degree of node $k+1$, we can write (5.23) as

$$\begin{aligned} k &= k\frac{\beta_C\beta_L}{2\psi - \beta_C\beta_L} \\ 1 &= \frac{\beta_C\beta_L}{2\psi - \beta_C\beta_L} + (k-1)\frac{\beta_L^2}{2\psi - \beta_L^2}\,. \end{aligned}$$

The first line requires that $\frac{\beta_C\beta_L}{2\psi - \beta_C\beta_L} = 1$. In conjunction with the second line, this

implies that $\beta_L = 0$, which in turn contradicts the first line unless $\beta_C = 0$ as well. We conclude that no solution to (5.23) exists which respects the symmetries of the star graph. On the other hand, simply adding a second copy of the star graph is sufficient introduce a solution. For example, in the union of two 5-stars, the algorithm we develop below to solve (5.23) finds that $\beta_C \approx 3.40$ and $\beta_L \approx 0.93$, with mean-square error below machine precision. In light of these examples, the conditions such that $\hat{\boldsymbol{\beta}}$ exists constitutes an interesting direction for future research.

5.4.1 Properties of $\hat{\boldsymbol{\beta}}$

While existence remains an open question, it is possible to provide a qualified uniqueness guarantee for (5.23). We will also prove several simple results about the "shape" of the entries of $\hat{\boldsymbol{\beta}}$ as functions of the entries of $\mathbf{d}$. Throughout this section, we assume that $\hat{\boldsymbol{\beta}}$ is sorted, so that $\hat{\beta}_1 \leq \hat{\beta}_2 \cdots \leq \hat{\beta}_n$.

Definition 23. *A vector $\boldsymbol{\beta}$ is* physical *if $\mathbf{e} \leq \boldsymbol{\beta} \leq (n-1)\mathbf{e}$ entrywise. A vector $\boldsymbol{\beta}$ is* well-behaved with parameter $\delta > 0$ *if, in addition, $\beta_n^2 \leq \mathbf{e}^T\boldsymbol{\beta} - \delta$.*

The bounds imposed by the physicality condition are in fact obeyed by the true expected collapsed degree vector $\mathbb{E}_\eta[\underline{]}$, provided that $\mathbf{d} \geq \mathbf{e}$ entrywise. Well-behavedness with parameter $\delta > 0$ is sufficient, but not necessary, to ensure that $\hat{\omega}_{ij} = f_{ij}(\hat{\boldsymbol{\beta}})(1 - f_{ij}(\hat{\boldsymbol{\beta}}))^{-1} > 0$ for all i and j. Let $\mathcal{B}_\delta$ denote the set of all physical, well-behaved vectors of (implied) fixed size n with a fixed parameter $\delta > 0$. Throughout, we will assume that δ is "sufficiently small;" this will not pose problems due to the inclusion $\mathcal{B}_{\delta'} \subset \mathcal{B}_\delta$ whenever $\delta' < \delta$. By construction, the function $\mathbf{h}$ defined by (5.22) is continuous, and indeed smooth, on $\mathcal{B}_\delta$.

We will show that (5.23) possesses at most one solution on $\mathcal{B}_\delta$. Let

$$\mathcal{L}(\boldsymbol{\beta}) = \|\mathbf{h}(\boldsymbol{\beta}) - \mathbf{d}\|_2^2 \tag{5.24}$$

be the square error associated with approximating $\mathbf{d}$ by $\mathbf{h}(\boldsymbol{\beta})$. Then, the problem

$$\min_{\boldsymbol{\beta}\in\mathcal{B}_\delta} \mathcal{L}(\boldsymbol{\beta}) \tag{5.25}$$

achieves its minimum value of 0 at the solutions of (5.23) in $\mathcal{B}_\delta$, provided there are any. We will show that (5.25) possesses at most one such minimum.

Our proof relies on an elementary form of the Mountain Pass Theorem [9], given as Lemma 6.1 in [31]. A closely related statement is given as Theorem 5.2 in [107].

Definition 24 (Palais-Smale Condition, [31])**.** *Let $q : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Let $\{\mathbf{a}_n\}$ be a sequence of points in $\mathbb{R}^n$ such that $q(\mathbf{a}_n)$ is bounded and $\|\nabla q(\mathbf{a}_n)\| \to 0$. The function q* satisfies the Palais-Smale condition *if any such $\{\mathbf{a}_n\}$ possesses a convergent subsequence.*

Theorem 11 (Mountain Pass Lemma in $\mathbb{R}^n$, [31, 9])**.** *Suppose that function $q : \mathbb{R}^n \to \mathbb{R}$ satisfies the Palais-Smale condition. Suppose further that:*

1. *$q(\mathbf{a}_0) = 0$.*
2. *There exists an $r > 0$ and $\alpha > 0$ such that $q(\mathbf{a}) \geq \alpha$ for all $\mathbf{a}$ with $\|\mathbf{a} - \mathbf{a_0}\| = r$.*
3. *There exists $\mathbf{a}'$ such that $\|\mathbf{a}' - \mathbf{a}_0\| > r$ and $q(\mathbf{a}') \leq 0$.*

Then, q possesses a critical point $\tilde{\mathbf{a}}$ with $q(\tilde{\mathbf{a}}) \geq \alpha$.

Our strategy is as follows. We will first show that all critical points of $\mathcal{L}$ are solutions of (5.23). We will then show that all such critical points are, furthermore, isolated local minima of (5.25). The existence of two such isolated local minima would trigger Theorem 11, implying the existence of an additional critical point with $\mathcal{L}(\boldsymbol{\beta}) > 0$. Since this is a contradiction, we will conclude that only one such minimum exists.

Our first step is to lower-bound the eigenvalues of the Jacobian $\mathbf{J}$ matrix of $\mathbf{h}$ at an arbitrary point $\boldsymbol{\beta}$. This Jacobian may be written

$$\mathbf{J} = (\mathbf{S} + \mathbf{D})\left(\mathbf{B}^{-1} - \frac{1}{4\psi}\mathbf{E}\right). \tag{5.26}$$

In this expression, $\mathbf{S}$ is the matrix with entries

$$s_{ij} = \begin{cases} \frac{f_{ij}(\boldsymbol{\beta})}{(1-f_{ij}(\boldsymbol{\beta}))^2} & i \neq j \\ 0 & i = j \, . \end{cases}$$

We have also defined $\mathbf{D} = \mathrm{diag}\mathbf{Se}$, and $\mathbf{B} = \mathrm{diag}(\boldsymbol{\beta})$. We note as a point of curiosity that $s_{ij} \approx \mathrm{var}\,(w_{ij})$ by (5.21), although our results here do not depend on this relationship. A derivation of (5.26) is supplied in the Supplementary Information. Let $\lambda_i(\mathbf{M})$ denote the ith eigenvalue of the matrix $\mathbf{M}$, sorted in ascending order. Thus, $\lambda_1(\mathbf{M})$ is the smallest eigenvalue of $\mathbf{M}$, and $\lambda_n(\mathbf{M})$ the largest.

Lemma 5. *Assume $n \geq 5$. Then,*

$$\lambda_1(\mathbf{J}) \geq \frac{1}{n(n-1)} \left(1 - \frac{2}{\sqrt{5}}\right) > 0 \, . \tag{5.27}$$

In particular, $\mathbf{J}$ is positive-definite and its eigenvalues are bounded away from zero on $\mathcal{B}_\delta$.

A proof is given in the Supplementary Information.

Lemma 6. *If $n \geq 5$ and $\boldsymbol{\beta}$ is a critical point of $\mathcal{L}$, then*

(a) $\boldsymbol{\beta}$ solves (5.23).

(b) The Hessian $\mathbf{H}$ of $\mathcal{L}$ at $\boldsymbol{\beta}$ is positive-definite.

Proof. To prove (a), we compute the gradient of $\mathcal{L}$:

$$\nabla \mathcal{L}(\boldsymbol{\beta}) = 2(\mathbf{h}(\boldsymbol{\beta}) - \mathbf{d})^T \mathbf{J}(\boldsymbol{\beta}) \, . \tag{5.28}$$

By Lemma 5, $\mathbf{J}(\boldsymbol{\beta})$ is positive-definite and therefore full-rank on $\mathcal{B}_\delta$. It follows that $\nabla \mathcal{L}(\boldsymbol{\beta}) = 0$ iff $\mathbf{h}(\boldsymbol{\beta}) = \mathbf{d}$, or, equivalently, iff $\mathcal{L}(\boldsymbol{\beta}) = 0$.

To prove (b), we calculate the entries of the Hessian. These are

$$\mathbf{H}(\boldsymbol{\beta})_{ij} = 2 \sum_{\ell=1}^{n} \left[(h_\ell(\boldsymbol{\beta}) - d_\ell) \frac{\partial^2 h_\ell}{\partial \beta_i \partial \beta_j} + \frac{\partial h_\ell}{\partial \beta_i} \frac{\partial h_\ell}{\partial \beta_j} \right] .$$

The first term vanishes at critical points. Recognizing the second as an outer product of the rows of $\mathbf{J}$, we can write the Hessian at critical points as

$$\mathbf{H}(\boldsymbol{\beta}) = 2\sum_{\ell=1}^{n} \mathbf{J}_\ell(\boldsymbol{\beta})\mathbf{J}_\ell(\boldsymbol{\beta})^T.$$

Since $\mathbf{J}$ is full rank by Lemma 5, the sum is full rank and therefore positive-definite. This completes the proof. □

We immediately obtain:

Corrolary 7. *If $n \geq 5$, then each critical point of $\mathcal{L}$ is an isolated local minimum, and there are finitely many of them.*

For the second clause, we rely on the fact that $\mathcal{B}_\delta$ is closed and bounded.

Lemma 7. *If $n \geq 5$, the restriction of $\mathbf{h}$ to $\mathcal{B}_\delta$ satisfies the Palais-Smale condition.*

Proof. Taking norms in (5.28) and lower-bounding the righthand side, we obtain

$$\|\nabla\mathcal{L}(\boldsymbol{\beta})\|_2 \geq \lambda_1\left(\mathbf{J}(\boldsymbol{\beta})\right)\|h(\boldsymbol{\beta}) - \mathbf{d}\|_2 \ .$$

Since $\lambda_1(\mathbf{J}(\boldsymbol{\beta}))$ is bounded away from zero on $\mathcal{B}_\delta$ by Lemma 5, the only sequences $\{\boldsymbol{\beta}_t\}$ in $\mathcal{B}_\delta$ that satisfy $\|\nabla\mathcal{L}(\boldsymbol{\beta}_t)\|_2 \to 0$ must also satisfy $\mathbf{h}(\boldsymbol{\beta}_t) \to \mathbf{d}$. By Lemma 6, there are finitely many solutions to (5.23), and therefore any such sequence has a finite number of limit points. The sequence $\{\boldsymbol{\beta}_t\}$ then possesses a subsequence that converge to each of these limit points, which completes the proof. □

Theorem 12. *If $n \geq 5$, there exists at most one $\hat{\boldsymbol{\beta}}$ in the set $\mathcal{B}_\delta$ such that $\mathbf{h}(\hat{\boldsymbol{\beta}}) = \mathbf{d}$.*

Proof. Suppose that there were two solutions $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ in $\mathcal{B}_\delta$. Since $\mathcal{L}$ satisfies the Palais-Smale condition (Lemma 7), we check conditions (1)-(3) of Theorem 11 are satisfied. Condition (1) requires that $\mathcal{L}(\boldsymbol{\beta}_0) = 0$, which is true by hypothesis. Condition (2) requires that there exists $r > 0$ and $\alpha > 0$ such that $h(\boldsymbol{\beta}) \geq \alpha$ for all $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = r$. This follows from Taylor-expanding $\mathcal{L}$ around $\boldsymbol{\beta}_0$ and using the positive-definiteness of $\mathbf{H}$. Applying Lemma 6 yields the existence of such an r and

α, and further implies that r may be taken to be arbitrarily small. In particular, r may be taken to be smaller than $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1\|$, which in turn supplies condition (3). Applying Theorem 11, we conclude that there exists a critical point $\tilde{\boldsymbol{\beta}}$ of $\mathcal{L}$ such that $\mathcal{L}(\tilde{\boldsymbol{\beta}}) \geq \alpha > 0$. But this contradicts Lemma 6. We conclude that at most one solution to (5.23) exists in $\mathcal{B}_\delta$, as was to be shown. □

Numerical experiments suggest that the solution to (5.23), if it exists, may be unique in the positive orthant $\mathbb{R}^n_+$. If true, this would be a stronger result than that provided by Theorem 12, which requires physicality and well-behavedness. Extending Theorem 12 to cover the full nonnegative orthant would be an interesting direction of future work.

The following theorem specifies several properties of $\hat{\boldsymbol{\beta}}$, provided that it exists.

Theorem 13. *Let $n \geq 5$. Suppose that $\hat{\boldsymbol{\beta}} \in \mathcal{B}_\delta$ solves (5.23). Then,*

(a) The map $d_i \mapsto \hat{\beta}_i$ is nondecreasing.

(b) Furthermore, $\hat{\beta}_i - \hat{\beta}_j \leq d_i - d_j$.

(c) Finally, $\hat{\boldsymbol{\beta}} \leq \mathbf{d}$ entrywise.

A proof of this result is given in the Supplementary Information.

5.4.2 Algorithms

Having proven some properties of the solutions of (5.23), it remains to develop an algorithm to find these solutions. While it is possible to use standard gradient-based methods, this task is complicated by the ill-conditioned Jacobian of $\mathbf{h}$. Ill-conditioning arises from dramatic heterogeneity in the entries of $\mathbf{S}$. For example, in the experiments shown in Figure 5-2 in the next section, the observed and estimated values of σ_{ij} span four orders of magnitude, implying that entries of $s_{ij} \approx \sigma_{ij}^2$ span roughly eight. Because of this, methods based on the full Jacobian, such as standard implementations of gradient descent or Newton's method, may require impractically small step-sizes in order to avoid pathological behavior.

We instead adopt a coordinate-wise approach. Suppose we have a current estimate $\hat{\boldsymbol{\beta}}^{(t-1)}$. We obtain an estimate of $\mathbf{d}$ given by $\hat{\mathbf{d}}^{(t-1)} = \mathbf{h}(\hat{\boldsymbol{\beta}}^{(t-1)})$. To update the ith coordinate of $\hat{\boldsymbol{\beta}}^{(t-1)}$, we hold all other $n-1$ coordinates fixed, and define $\hat{\beta}_i^{(t)}$ to be the value of b that solves the equation

$$h_i\left(\hat{\beta}_1^{(t-1)}, \ldots, \hat{\beta}_{i-1}^{(t-1)}, b, \hat{\beta}_{i+1}^{(t-1)}, \ldots, \hat{\beta}_n^{(t-1)}\right) = d_i \,. \tag{5.29}$$

We repeat this process for each of the n coordinates, obtaining a fully updated new estimate $\hat{\boldsymbol{\beta}}^{(t)}$. We iterate this sweep over the coordinates until a desired error function drops below a user-specified tolerance. A standard choice for the error function is the mean-square error $n^{-1}\mathcal{L}(\boldsymbol{\beta})$, with $\mathcal{L}$ as in (5.24). Algorithm 6 formalizes the solution method. The call Solve_b solves a single-variable equation for b. In the accompanying code (see "Software"), we implement Solve_b with options to use either the `root_scalar()` function supplied by `python`'s `scipy` package or a bespoke Newton-type method.

Algorithm 6: Computation of $\hat{\beta}$

Input: degree sequence $\mathbf{d} \in \mathbb{Z}_+^n$, initial guess $\hat{\boldsymbol{\beta}}^{(0)} \in \mathbb{R}_+^n$, tolerance ϵ

Initialization: $t \leftarrow 0$, $\gamma \leftarrow \infty$

while $\gamma^{(t)} > \epsilon$ **do**

  for $i = 1, \ldots, n$ **do**

    $\hat{\beta}_i^{(t)} \leftarrow \text{Solve}_b\{h_i(\hat{\beta}_1^{(t-1)}, \ldots, \hat{\beta}_{i-1}^{(t-1)}, b, \hat{\beta}_{i+1}^{(t-1)} \ldots, \hat{\beta}_n^{(t-1)}) = d_i\}$

  $\gamma^{(t)} \leftarrow n^{-1}\mathcal{L}(\hat{\boldsymbol{\beta}}^{(t)})$

  $t \leftarrow t + 1$

Output: $\hat{\boldsymbol{\beta}}^{(t)}$

In order to ensure that this algorithm is well-defined, we will show that the update given by (5.29) possesses a unique solution under mild conditions.

Lemma 8. *Assume that* $\mathbf{d} > \mathbf{0}$ *and* $\boldsymbol{\beta}^{(t-1)} > \mathbf{0}$ *entrywise. Then, for each* i, (5.29) *possesses a unique solution in* b *on the open interval* $\left(0, \frac{2\psi^{(t-1)}}{\max_{\ell \neq i} \beta_\ell^{(t-1)}}\right)$.

Remark. The hypotheses of Lemma 8 can be ensured by removing degree-zero nodes from $\mathbf{d}$ and initializing $\boldsymbol{\beta}^{(0)} > \mathbf{0}$.

Proof. To prove existence, we note that h_i is a continuous function of β_i. We have $h_i(\beta_1, \ldots, 0, \ldots \beta_n) = 0$ and

$$\lim_{\beta \to \frac{2\psi^{(t-1)}}{\max_{\ell \neq i} \beta_\ell}} h_i(\beta_1, \ldots, \beta, \ldots, \beta_n) = \infty \, .$$

The Intermediate Value Theorem then provides existence.

To show uniqueness, it suffices to check the derivative (cf. (5.26))

$$\frac{\partial h_i(\boldsymbol{\beta})}{\partial \beta_i} = \left(\frac{1}{\beta_i} - \frac{1}{2\psi} \right) \sum_{\ell \neq i} \frac{f_{i\ell}(\boldsymbol{\beta})}{(1 - f_{i\ell}(\boldsymbol{\beta}))^2} \, .$$

When $\boldsymbol{\beta} > 0$, this expression is strictly positive. The function h_i is therefore strictly increasing on I, proving uniqueness. □

While we have existence, uniqueness, and convergence guarantees for each coordinate update, we possess no such guarantees for Algorithm 6 as a whole. Additionally, it may be the case that some elements of the sequence $\{\hat{\boldsymbol{\beta}}^{(t)}\}$ produce estimates $\hat{\omega}^{(t)}$ of the adjacency matrix in which some entries are negative. However, we have never observed Algorithm 6 to fail to converge to a solution in which all entries of $\hat{\omega}^{(t)}$ are positive. Additionally, when a solution to (5.23) exists in $\mathcal{B}_\delta$ for some δ, we have never observed Algorithm 6 to fail to find this solution. In practice, an analyst can assess the success of the algorithm by checking that (a) the mean-square error is near zero and that (b) the corresponding estimate $\hat{\boldsymbol{\Omega}}$ has nonnegative entries. Both such checks are implemented in the accompanying software.

5.5 Experiments

In this section, we describe a sequence of experiments exploring the behavior of Algorithm 6; the accuracy of the estimator $\hat{\boldsymbol{\beta}}$; the disparity between $\hat{\boldsymbol{\Omega}}^0$ and $\hat{\boldsymbol{\Omega}}^1$ on empirical networks; and implications for downstream tasks such as modularity maximization.

5.5.1 Synthetic Data

To study the convergence behavior of Algorithm 6, we test it on two synthetic degree sequences. The "uniform" sequence consists of 200 independent copies of $2(u+1)$, where u is a discrete uniform random variable on the interval $[0, 50]$. We contrast this with a "Zipf" sequence $\mathbf{d}_2$, generated by sampling 200 copies of $2z$, where z is distributed as a Zipf random variable with parameter $\alpha = 2$. These degree sequences are shown in Figure 5-1(a). By design, the uniform sequence is relatively homogeneous in its degrees, while the Zipf sequence possesses a small number of extremely high-degree nodes.

We then estimated $\hat{\boldsymbol{\beta}}$ for each of these degree sequences using Algorithm 6, initialized with $\boldsymbol{\beta}^{(0)} = \mathbf{e}$. The estimates for the uniform sequence $\mathbf{d}_1$ converge rapidly, as shown in panel (b), and after two rounds the iterates cannot be distinguished by eye from the final estimate. The final estimate $\hat{\boldsymbol{\beta}}$ is both physical and well-behaved. By Theorem 12, it is the only such solution to (5.23). In contrast, the estimates for the Zipf-distributed sequence $\mathbf{d}_2$, shown in panel (c), require many rounds to converge. Figure 5-1(d) compares the differing convergence rates. The vertical axis gives the mean-square error (MSE) $\frac{1}{n}\left\|\mathbf{h}(\boldsymbol{\beta}) - \mathbf{d}\right\|_2$. While the MSE for the uniform degree sequence converges to within machine precision after 14 rounds, the Zipf iterates require over 200 iterations to reach an MSE below 10^{-6}. The resulting estimate $\hat{\boldsymbol{\beta}}$ is physical but not well-behaved, and Theorem 12 is therefore insufficient to provide a uniqueness guarantee.

Algorithm 6 also allows us to perform some bootstrap-style tests of Conjectures 1. Recall that this conjecture asserts that the constant $u(\mathbf{d})$, which bounds the effect of perturbations of $\mathbf{d}$ on β, is no larger than 1. The size of $u(\mathbf{d})$ in turn influences the tightness of the error bounds derived in Section 5.3. Figure 5-1(e-f) shows the results of a simple experiment in which we use our estimator $\hat{\boldsymbol{\beta}}$ as a surrogate for $\boldsymbol{\beta}$. For each degree sequence, we repeatedly sample i and j from $\binom{[n]}{2}$. We then compute $\hat{\boldsymbol{\beta}}' = \hat{\boldsymbol{\beta}}(\mathbf{d} + \mathbf{e}_i + \mathbf{e}_j)$ and compare it to $\hat{\boldsymbol{\beta}}$. Filled dots show the maximum absolute change, $\left\|\hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}}\right\|_\infty$, while empty dots give the mean absolute change $\frac{1}{n}\left\|\hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}}\right\|_1$. Un-

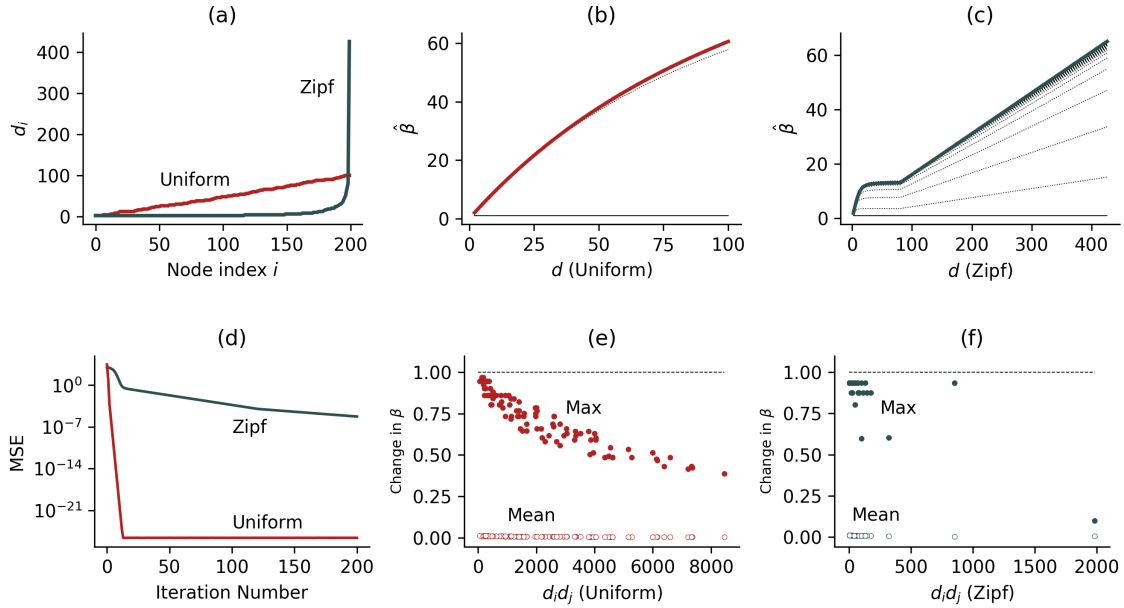


Figure 5-1: Experiments on synthetic data.
(a): The uniform and Zipf degree sequences described in the text. (b): Iterates of Algorithm 6 for the uniform degree sequence. The final iterate is highlighted. (c): Iterates of Algorithm 6 for the Zipf degree sequence. The final iterate is highlighted. (d): Mean-square error (MSE) in Algorithm 6 for the uniform (red) and Zipf (gray) degree distributions as a function of the iteration number. (e): Bootstrap estimates of $\|\boldsymbol{\beta}(\mathbf{d}+\mathbf{e}_i+\mathbf{e}_j)-\boldsymbol{\beta}\|_\infty$ (filled points) and $n^{-1}\|\boldsymbol{\beta}(\mathbf{d}+\mathbf{e}_i+\mathbf{e}_j)-\boldsymbol{\beta}\|_1$ (empty circles)for the uniform degree sequence. Each point corresponds to a uniformly random choice of distinct indices i and j. (f): As in (e), for the Zipf degree sequence.

der Conjectures 1, we would expect that $\left\|\hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}}\right\|_\infty \leq 1$, which is indeed the case for both degree sequences. These results may be viewed as heuristic supports of Conjectures 1. Additionally, $\frac{1}{n}\left\|\hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}}\right\|_1 \ll 1$. This observation suggests the possibility of substantially tightening the error bounds given in Section 5.3 by controlling the ℓ^1-norm rather than the ℓ^∞ norm, an interesting problem which we leave to future work.

5.5.2 Evaluation on an Empirical Contact Network

Our evaluation data set is a contact network among students in a French secondary school, called `contact-high-school` [134, 25]. During data collection, each student wore a proximity sensor. An interaction between two students was logged by their respective sensors when the students were face-to-face and within approximately 1.5m of each other. Edges are time-stamped, although we do not use any temporal information the present experiments. The original data set contains $n = 327$ nodes and $m = 189,928$ distinct interactions.

We first test the accuracy of the estimator $\hat{\boldsymbol{\Omega}}^1$, using $\hat{\boldsymbol{\Omega}}^{\text{mc}}$ as a reliable estimate of the true mean $\boldsymbol{\Omega}$. Because of the scaling issues associated with estimating $\hat{\boldsymbol{\Omega}}^{\text{mc}}$ on $m \approx 2 \times 10^5$ edges, we constructed a data subset based on a temporal threshold τ, chosen to incorporate approximately the last 5% of the original interaction volume. The resulting subnetwork has 268 nodes and 10,026 edges. To estimate the ground-truth moments $\eta_{\mathbf{d}}$, we estimated $\hat{\boldsymbol{\Omega}}^{\text{mc}}$ on the subnetwork from 10^7 samples at intervals of 10^3 steps. This computation required approximately one week on a single thread of a modern server.

In Figure 5-2(a)-(b), we show the distributions of degrees and entries of $\mathbf{w}$ for this subnetwork. Figure 5-2(a) depicts the heterogoenous degree distribution, with standard deviation larger than the average degree. While most nodes have small degrees, there are twelve whose degree exceeds n. Figure 5-2(b) shows the clumping of edges between pairs of nodes. On average, two students who interact at all interact nearly ten times, but there is substantial deviation around this average. Almost half

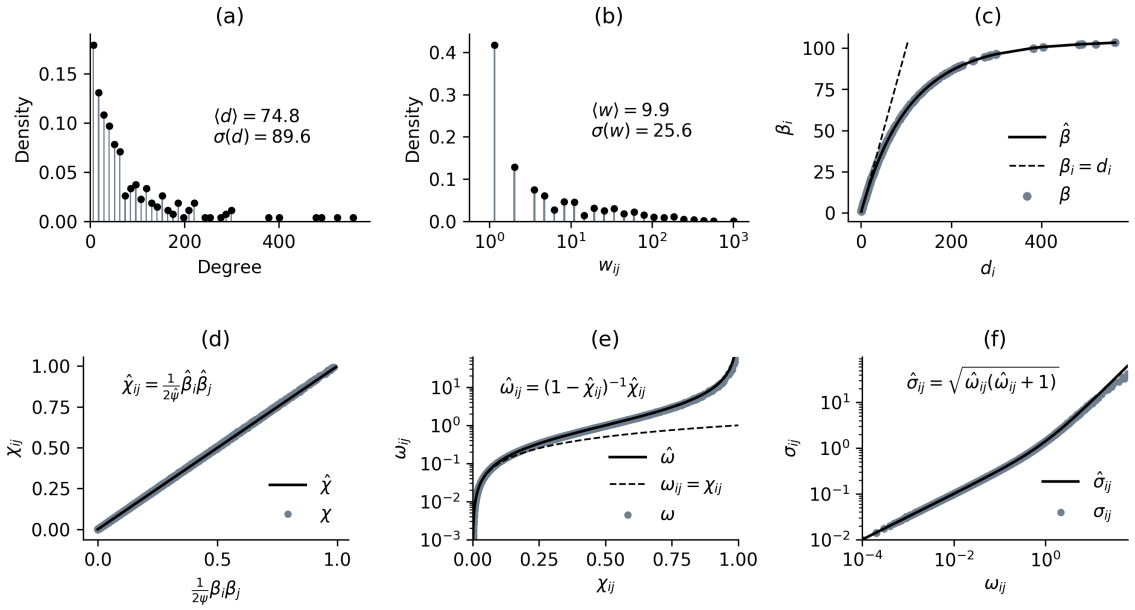


Figure 5-2: Experiments on the `contact-high-school` network subset. (a): Degree distribution of the subnetwork. The mean degree $\langle d \rangle$ and standard deviation of the degree $\sigma(d)$ are shown. (b): Distribution of the entries of $\mathbf{w}$. Note the logarithmic horizontal axis. (c): Collapsed degree sequence $\hat{\boldsymbol{\beta}}$ learned from $\mathbf{d}$ via Algorithm 6. Dashes give the line of equality. (d): Approximation of $\boldsymbol{\chi}$ via (5.18). (e): Approximation of $\boldsymbol{\Omega}$ via (5.20). Note the logarithmic vertical axis. Dashes give the line of equality. (f): Approximation of $\sigma_{ij} = \sigma(W_{ij})$ via (5.21). Note the log-log axis. In (c)-(f), simulated moments (gray dots) are obtained via the Monte Carlo estimator using Algorithm 5; see main text for details.

of all pairs interact just once. In contrast, a small number of pairs interact over 100 times, and one over 1,000.

In Figure 5-2(c)-(f), we show the construction of estimators for the moments of $\mathbf{\Omega}$ under the uniform random graph model with the observed degree sequence $\mathbf{d}$. In Figure 5-2(c), the solid line shows the estimate $\hat{\boldsymbol{\beta}}$ output by Algorithm 6, plotted against the degree sequence. Points give the MCMC estimate for $\boldsymbol{\beta}$. The agreement is almost exact. The estimate $\hat{\boldsymbol{\beta}}$ is physical and well-behaved. Theorem 12 implies that it is the only physical, well-behaved solution to (5.23).

In Figure 5-2(d), we estimate $\hat{\chi}_{ij} \approx f_{ij}(\hat{\boldsymbol{\beta}}) = \frac{\hat{\beta}_i \hat{\beta}_j}{2\hat{\psi}}$, again finding the agreement to be near exact. In (e), we estimate $\hat{\omega}_{ij} \approx (1 - \hat{\chi}_{ij})^{-1} \hat{\chi}_{ij}$. The agreement with data is again excellent, although there is a small amount of visible overestimation of ω_{ij} when χ_{ij} is large. Finally, (f) uses (5.21) to compute an estimator $\hat{\sigma}_{ij} = \sqrt{\hat{\omega}_{ij}(\hat{\omega}_{ij} + 1)}$ of σ_{ij} the standard deviation of W_{ij}. The agreement is strong through roughly $\omega_{ij} \approx 10$, and begins to overestimate σ_{ij} for larger values.

Figure 5-2(c) and (e) also highlight the relationship of $\hat{\mathbf{\Omega}}^1$ and $\hat{\mathbf{\Omega}}^0$. The dashed lines in these figures represent two linear approximations that can be made to yield the latter from the former. First, we approximate $\boldsymbol{\beta} = \mathbf{d}$ (dashed line, Figure 5-2(c)). This approximation holds good when d_i is small, since then the number of parallel edges incident to node i should be small – i.e. $W_{ij} \approx X_{ij}$. Then, we approximate $\mathbf{\Omega} = \boldsymbol{\chi}$ (dashed line, Figure 5-2(e)). This approximation should hold for small entries of $\mathbf{\Omega}$, since in this case χ_{ij} is small and $(1 - \chi_{ij})^{-1} \approx 1$. As the plots indicate, these approximations are indeed accurate when d_i and ω_{ij} are small. These conditions correspond roughly to the "large, sparse" heuristics used frequently in the literature. We can therefore view $\hat{\mathbf{\Omega}}^0$ as a first-order approximation to $\hat{\mathbf{\Omega}}^1$ near the large, sparse regime. Conversely, we can view $\hat{\mathbf{\Omega}}^1$ as a nonlinear correction to $\hat{\mathbf{\Omega}}^0$ as we depart from that regime.

Figure 5-3 compares the overall performance of the estimators $\hat{\mathbf{\Omega}}^0$ and $\hat{\mathbf{\Omega}}^1$. We compute the entrywise relative error $\mathcal{E}_{ij}(\hat{\mathbf{\Omega}}) = (\hat{\omega}_{ij}^{\text{mc}})^{-1}(\hat{\omega}_{ij} - \hat{\omega}_{ij}^{\text{mc}})$ when approximating $\hat{\mathbf{\Omega}}^{\text{mc}} \approx \mathbf{\Omega}$ with both methods. Cells are shaded according to the magnitude and sign of the error. The CL estimator in (a) displays systematic bias, underestimating the

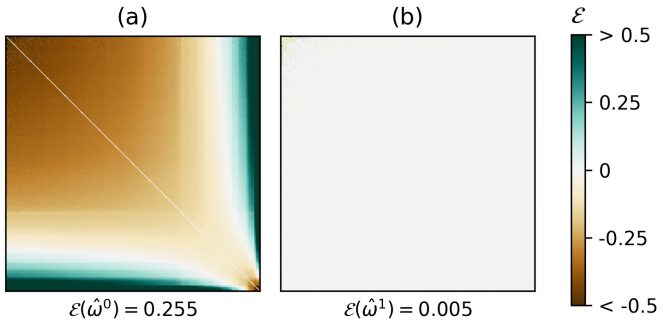


Figure 5-3: Relative error when approximating $\mathbf{\Omega}$ under the uniform model for the `contact-high-school` subnetwork.
Shading gives the relative error for approximating $\mathbf{\Omega}$ under the uniform model for the `contact-high-school` subnetwork. (a): Using the CL estimator $\hat{\mathbf{\Omega}}^0$. (b): Using the present estimator $\hat{\mathbf{\Omega}}^1$. Node degrees in each matrix increase left to right and top to bottom. The "ground truth" is provided by $\hat{\mathbf{\Omega}}^{\text{mc}}$, computed as in Figure 5-2.

density of edges between nodes of similar degrees and overestimating the density of edges between nodes with highly disparate degrees. The mean absolute relative error of the Chung-Lu estimate is $\mathcal{E}(\hat{\mathbf{\Omega}}^0) = \binom{n}{2}^{-1} \sum_{ij} \left|\mathcal{E}_{ij}(\hat{\mathbf{\Omega}}^0)\right| \approx .255$, indicating that a typical entry of $\hat{\mathbf{\Omega}}^0$ is off by over 25%. In contrast, $\hat{\mathbf{\Omega}}^1$ evaluated in (b) has almost no visible bias. Some large residuals are visible for entries ω_{ij} in which both d_i and d_j are small (top left corner), although it is difficult to evaluate to what extent these residuals reflect error in $\hat{\mathbf{\Omega}}^1$ or in the challenge to $\hat{\mathbf{\Omega}}^{\text{mc}}$ to estimate these edge densities in finite runtime. The mean absolute relative error $\mathcal{E}(\hat{\mathbf{\Omega}}^1)$ is roughly 0.5%, an improvement over $\hat{\mathbf{\Omega}}^0$ of a full order and a half of magnitude.

5.5.3 Application: Modularity Maximization in Dense Contact Networks

Let $\ell : N \to \mathcal{L}$ be a function that assigns to node i a label $\ell_i \in \mathcal{L}$. The *modularity* of the partition ℓ with respect to matrix $\mathbf{w}$ and null model ρ is given by

$$Q(\ell; \rho) = \frac{1}{2m} \sum_{ij} \left[w_{ij} - \mathbb{E}_\rho[W_{ij}]\right] \mathbb{1}(\ell_i, \ell_j) \,. \tag{5.30}$$

The normalization by $2m$ ensures that $-1 \leq Q(\ell;\rho) \leq 1$. Intuitively, $Q(\ell;\rho)$ is high when nodes that are more densely connected than expected by chance (under the specified null) are grouped together. Maximizing this quantity with respect to ℓ may therefore be reasonably expected to identify modular ("community") structure in the network [143, 146]. Exact modularity maximization is NP-hard [39] and subject to theoretical limitations in networks with modules of heterogeneous sizes [75]. Despite this, it remains one of the most popular methods for practical community detection at scale [33].

In most implementations, ρ is not explicitly specified – rather, the expectation $\mathbb{E}_\rho[W_{ij}]$ is "hard-coded" as equal to $\hat{\omega}^0_{ij}$. From a statistical perspective, this reflects an implicit choice of ρ as the Chung-Lu model [51], which preserves expected degrees and indeed possesses the given first moment.[2] Modifications are possible; the best known is perhaps the resolution adjustment that replaces $\hat{\mathbf{\Omega}}^0$ with $\gamma\hat{\mathbf{\Omega}}^0$ for some $\gamma > 0$ [174]. Other adjustments may incorporate spatial structure [71] or adjust for the inclusion of self-loops in the null space [43]. When we wish to perform modularity maximization against a null that deterministically preserves degree sequences, however, $\hat{\mathbf{\Omega}}^0$ is at best an approximation. We expect this approximation to perform adequately for the configuration model (cf. Theorem 7), and very poorly for the uniform model (previous subsection). The Monte Carlo estimate $\hat{\mathbf{\Omega}}^{\text{mc}}$ can be used for very small data sets, but rapidly becomes computationally infeasible for larger ones. In these cases, we can use the present estimator $\hat{\mathbf{\Omega}}^1$ instead.

Recent work has highlighted the importance of studying the *modularity landscape*, especially the set of local maxima of Q, rather than restricting attention to a single partition. One reason for this is the phenomenon of *degeneracy* – in many practical contexts, a given network will possess many distinct local maxima with modularity comparable to the global maximum [90]. A second reason is model-specification. As shown in [144], maximization of Q is related to maximum-likelihood inference in a planted-partition stochastic blockmodel. When the planted-partition model is

[2] We note that alternative justifications of the use of $\hat{\mathbf{\Omega}}^0$ exist, including connections to the stability of Markov chains [62] and to stochastic block models [144].

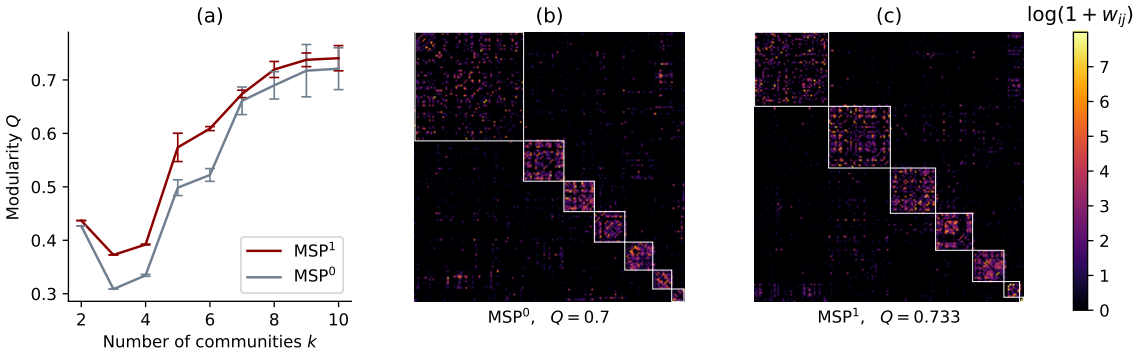


Figure 5-4: Modularity experiments on the full `contact-high-school` network. (a): Performance of MSP on the full `contact-high-school` network using the CL modularity matrix $\mathbf{M}^0$ and the approximate uniform modularity matrix $\mathbf{M}^1$, over 100 batches of 50 repetitions each. Solid lines give the average modularity, and error bars give two standard deviations from the mean. (b): Example partition using $\mathbf{M}^0$. (c): Example partition using $\mathbf{M}^1$. To generate (b) and (c), the best partition of 500 runs was chosen for each algorithm variant. Each run was initialized with $k = 8$; in the best partition, however, only 7 labels are actually used. Colors are shown on a log scale.

unrealistic as a generative model for the data, modularity-maximization amounts to inference in a mis-specified model. Degeneracy is a common symptom of this problem, but observing it requires locally optimizing Q multiple times. For these reasons among others, ensemble-based methods that implicitly average over local optima, such as those of [220], may be preferable. With these considerations in mind, our aim in this section is not to show that the use of $\hat{\mathbf{\Omega}}^1$ is strictly superior for this task when compared to $\hat{\mathbf{\Omega}}^0$ for either one-shot or ensemble-based modularity maximization. Rather, we will argue that the corresponding modularity landscapes are significantly different on data sets of practical interest, and that it is therefore methodologically unsafe to interchange these estimators without carefully scrutinizing the results.

Our first data set for this experiment is the full `contact-high-school` network, consisting of $n = 327$ nodes and $m = 189,928$ edges as described in the previous subsection. The computation of $\hat{\mathbf{\Omega}}^{\mathrm{mc}}$ is indeed infeasible for a graph this dense, and we therefore use $\hat{\mathbf{\Omega}}^1$ as an estimate. This setting highlights the utility of $\hat{\mathbf{\Omega}}^1$, since otherwise we would have no practical way to compute the uniform expectation.

To approximately maximize (5.30), we employ the multiway spectral partitioning

(MSP) algorithm of [221], which generalizes the spectral bipartitioning algorithm of [143]. While greedy methods often enjoy superior performance [33], spectral methods have the advantage of depending strongly on the structure of the observed graph and the null model employed, and are relatively insensitive to choices made during the runtime of the algorithm. Spectral methods are therefore ideal for highlighting differences in the modularity landscapes induced by alternative null models. The algorithm requires the analyst to specify a null model and a desired number of communities k. The core of the approach is to use a low-rank approximation of the *modularity matrix*, $\mathbf{M} = \mathbf{w} - \mathbb{E}_\rho[\mathbf{W}]$. This approximation induces a map from the vertices of G to a low-dimensional vector space. Vectors in this space are clustered according to their relative angles using a procedure reminiscent of k-means to produce the community assignment. Because the clustering algorithm involves a stochastic starting condition, it is useful to run the algorithm multiple times and choose the highest modularity partition from among the repetitions. We refer the reader to [221] for details, and to the code accompanying this paper for an implementation of MSP for arbitrary modularity matrices (see "Software").

We ran this algorithm using both the CL modularity matrix $\mathbf{M}^0 = \mathbf{w} - \hat{\mathbf{\Omega}}^0$ and the approximate uniform modularity matrix $\mathbf{M}^1 = \mathbf{w} - \hat{\mathbf{\Omega}}^1$. We refer to these two algorithmic variants as MSP^0 and MSP^1, respectively. Since $\hat{\mathbf{\Omega}}^0$ and $\hat{\mathbf{\Omega}}^1$ produce very different null matrices, the modularity matrices $\mathbf{M}^0$ and $\mathbf{M}^1$ are themselves very different – the mean absolute relative error of using the latter to estimate the former is approximately 32%. We would therefore expect MSP^0 and MSP^1 to behave very differently in this task. We allowed the number of communities k to vary between 2 and 10. For each value of k, we ran MSP^0 and MSP^1 in 100 batches of 50 repetitions. From each batch of 50, the highest-modularity partition was chosen, resulting in 100 partitions per value of k. Figure 5-4(a) shows that MSP^1 tends to find higher modularity partitions than MSP^0 on this data set. The difference is especially large when k is small, but a substantial difference between the means is noticeable even for larger values. While partitions under $\mathbf{M}^0$ exist that are comparable to those under $\mathbf{M}^1$, it appears to be more difficult for MSP^0 to find them. Panels (b) and

(c) shed some light on the differing behavior of the two algorithms. Partitions under MSP^0 tends to display a larger, less cohesive community ((b), top left) alongside smaller, more tightly interconnected ones. Partitions under MSP^1 (c) tend to display communities that are slightly more uniform in size.

It is reasonable to object that modularity values under MSP^1 and MSP^0 should not be compared, since these objectives are defined with respect to differing null matrices. In this specific case, the objection is not borne out numerically, however – "cross-evaluating" the partitions on the opposite matrices changes the modularities only minimally. Evaluating the MSP^0 partition in Figure 5-4(b) on the modularity matrix $\mathbf{M}^1$ gives $Q = 0.699$, while evaluating the MSP^1 partition on $\mathbf{M}^0$ yields $Q = 0.731$. On this data set, MSP^1 searches the energy landscape of MSP^0 more efficiently than does MSP^0 itself.

It should be noted that this behavior is data-set dependent. The opposite case occurs in the `contact-primary-school` network [194, 25], which used similar sensors to construct an interaction network among students in a French primary school. On this data, MSP^0 and MSP^1 perform similarly for $k \leq 6$ communities (Figure 5-5), with the former consistently outperforming the latter for $k \geq 7$. The illustrative partitions in panels (b) and (c) suggest MSP^1 tends to prefer partitions with fewer communities. Whereas MSP^0 chooses a partition with 7 communities, MSP^1 chooses one with just 5 (both having been initialized at $k = 8$). These illustrations emphasize that MSP^1 and MSP^0 explore different modularity landscapes; that the relative advantages of each algorithm depend on the data; and that the landscape for MSP^1 can be tractably computed under the methodology we have introduced here.

5.6 Discussion

Much existing network theory is explicitly designed for large, sparse data. However, many networks of interest are sufficiently dense to diverge significantly from the predictions of large, sparse theory. We have highlighted this phenomenon in the context of dense multigraphs, with a focus on estimating the expected adjacency matrix $\mathbf{\Omega}$

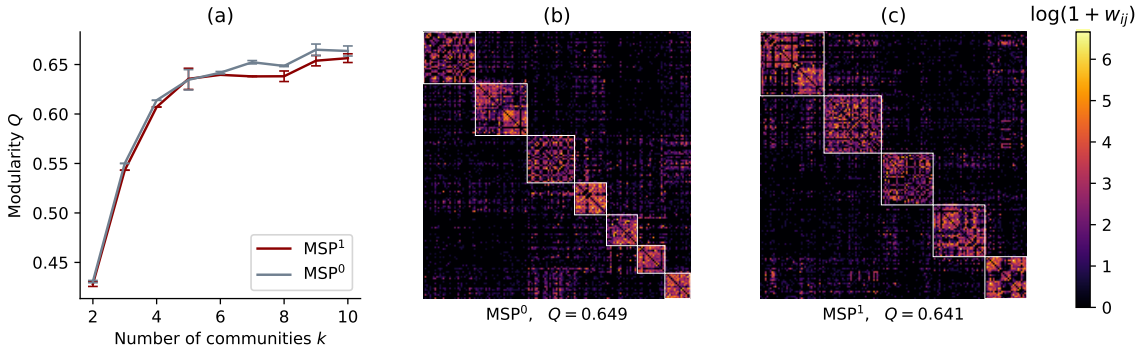


Figure 5-5: Modularity experiments on the full `contact-primary-school` network. This figure is in all methodological details identical to Figure 5-4, using the study data set `contact-primary-school` [195, 25]. In (b)-(c), both algorithms were initialized with $k = 8$.

of a random multigraph with specified degree sequence. We have shown that, rather than falling back to computationally expensive MCMC, we can construct an accurate estimator $\hat{\mathbf{\Omega}}^1$ using an indirect, dynamical approach. Use of this estimator can in turn have significant impact on the results of downstream data analyses.

There are several directions for future work on the moments of uniform random graphs with fixed degree sequences. As previously noted, the error bounds on $\hat{\boldsymbol{\chi}}$ and $\hat{\mathbf{\Omega}}^1$ derived in Section 5.3 appear quite loose when when compared against the empirical results in Figures 5-1 and 5-2. The derivation of tighter error bounds would be helpful for researchers seeking practical accuracy guarantees. Progress on this front appears to be hindered by the complex combinatorial structure of the space $\mathcal{G}_{\mathbf{d}}$; however, carefully chosen assumptions or approximations may allow headway. An additional avenue of exploration concerns the impact of the choice between $\hat{\mathbf{\Omega}}^1$ and $\hat{\mathbf{\Omega}}^0$ on downstream analyses. We have seen that the choice of null expectation can substantially change the performance of MSP, and that the direction of this effect depends on the data set. A better understanding of the properties of the data or algorithm that make certain estimators highlight better solutions would be most welcome.

We focused our attention on the derivation of an estimator for $\mathbf{\Omega}$. It may also be possible to derive expressions for higher moments using the same methodology.

Such moments would approximate expected densities of various motifs under the uniform model. Examples of interest may include wedge densities $\mathbb{E}[w_{ij}w_{jk}]$ and triangle densities $\mathbb{E}[w_{ij}w_{jk}w_{ik}]$. Parsing the stationarity conditions for these more complicated moments may be correspondingly more difficult. An alternative would be to construct mean-field estimates by computing the relevant statistics on $\hat{\mathbf{\Omega}}^1$ itself. An evaluation of the accuracy of this approach would potentially replace the need for computationally intensive MCMC sampling to estimate these quantities.

Finally, it may also be of interest to develop similar theory for uniform distributions over related spaces of graphs. For example, it would be possible to consider a uniform model including self-loops. The associated analysis would be nontrivial due to required modifications in the MCMC sampling procedure (see [76]), but one might reasonably hope to obtain parallel results. Directed graphs offer another important direction of generalization. It would be natural to define a uniform distribution over spaces of directed multigraphs with fixed in-degree and out-degree sequences. In this case, one might expect analysis to produce expressions for the moments of this distribution in terms of two collapsed degree sequences, corresponding to in- and out-degrees. These and other generalizations offer promising avenues of future work.

5.7 Additional Proofs

5.7.1 Proof of Theorem 10

Write

$$\mathbb{E}[w_{ij}\Delta_{ij}] = \mathbb{E}[w_{ij}(\Delta_{ij}^{+} - \Delta_{ij}^{-})] \tag{5.31}$$

$$= \frac{1}{z(\mathbf{W})}\mathbb{E}\left[w_{ij}\left(b_i b_j - 2y + b_i + b_j - \mathbf{x}_i^T\mathbf{x}_j - 2\right)\right] . \tag{5.32}$$

We have used the identity $w_{ij}x_{ij} = w_{ij}$ to eliminate instances of x_{ij} from this expression.

Starting with the first term, write

$$\begin{aligned}\mathbb{E}[w_{ij}b_ib_j] &= \omega_{ij}\mathbb{E}[b_ib_j]\left(1+\frac{\operatorname{cov}(w_{ij},b_ib_j)}{\omega_{ij}\mathbb{E}[b_ib_j]}\right)\\ &= \omega_{ij}\beta_i\beta_j\left(\frac{\mathbb{E}[b_ib_j]}{\beta_i\beta_j}+\frac{\operatorname{cov}(w_{ij},b_ib_j)}{\omega_{ij}\beta_i\beta_j}\right).\end{aligned}$$

Applying (5.13), the first term inside the parentheses satisfies

$$\left|\frac{\mathbb{E}[b_ib_j]}{\beta_i\beta_j}-1\right| \le \frac{2u_*}{\beta_i\vee\beta_j}.$$

To bound the second term, we apply Cauchy-Schwartz along with (5.13) and (5.14).

$$\begin{aligned}|\operatorname{cov}(w_{ij},b_ib_j)| &\le \sigma_{ij}\sqrt{\operatorname{var}(b_ib_j)}\\ &\le \sqrt{(\beta_i+6u_*)^2(\beta_j+6u_*)^2-(\beta_i\beta_j-2u_*(\beta_i\wedge\beta_j))^2}.\end{aligned}$$

We therefore obtain

$$|\mathbb{E}[w_{ij}b_ib_j]-\omega_{ij}\beta_i\beta_j| \le \frac{2u_*}{\beta_i\vee\beta_j}+\frac{\sigma_{ij}}{\omega_{ij}}\frac{\sqrt{(\beta_i+6u_*)^2(\beta_j+6u_*)^2-(\beta_i\beta_j-2u_*(\beta_i\wedge\beta_j))^2}}{\beta_i\beta_j}.$$

The remaining terms are simpler. We have

$$|\mathbb{E}[w_{ij}y]-\omega_{ij}\psi| = |\operatorname{cov}(w_{ij},y)| \le \sigma_{ij}\sqrt{\operatorname{var}(y)} \le \sigma_{ij}\sqrt{2v_*\psi},$$

where in the first inequality we have used Cauchy-Schwartz and in the second we have used (5.15).

To control the remaining terms, start by using the same argument to yield

$$|\mathbb{E}[w_{ij}b_i]-\omega_{ij}\beta_i| \le \sigma_{ij}\sqrt{2u_*\beta_i}.$$

Then, we obtain the upper bound

$$\begin{aligned}\mathbb{E}[w_{ij}(b_i+b_j)-\mathbf{x}_i^T\mathbf{x}_j-2] &\leq \mathbb{E}[w_{ij}(b_i+b_j)] \\ &\leq \omega_{ij}(\beta_i+\beta_j)+\sigma_{ij}\sqrt{2u_*}(\sqrt{\beta_i}+\sqrt{\beta_j})\,.\end{aligned}$$

Inserting our results into (5.32), we are able to write

$$\left|\mathbb{E}[w_{ij}\Delta_{ij}]-\frac{\omega_{ij}(\beta_i\beta_j-2\psi)}{z(\mathbf{W})}\right| \leq \frac{\epsilon'_{ij}(\boldsymbol{\beta})}{z(\mathbf{W})}\,, \tag{5.33}$$

where

$$\begin{aligned}\epsilon'_{ij}(\boldsymbol{\beta}) \triangleq \frac{2u_*}{\beta_i\vee\beta_j}+\frac{\sigma_{ij}}{\omega_{ij}}&\frac{\sqrt{(\beta_i+6u_*)^2(\beta_j+6u_*)^2-(\beta_i\beta_j-2u_*(\beta_i\wedge\beta_j))^2}}{\beta_i\beta_j} \\ &+\sigma_{ij}\sqrt{2v_*\psi_i}+\omega_{ij}(\beta_i+\beta_j)+\sigma_{ij}\sqrt{2u_*}(\sqrt{\beta_i}+\sqrt{\beta_j})]\,.\end{aligned}$$

From Lemma 4, we also have

$$\left|\delta^+_{ij}+\delta^-_{ij}-\frac{\beta_i\beta_j+2\psi\chi_{ij}}{z(\mathbf{W})}\right| \leq \frac{\epsilon^-_{ij}(\boldsymbol{\beta})+\epsilon^+_{ij}(\boldsymbol{\beta})}{z(\mathbf{W})}\,.$$

By definition, the righthand side is $\epsilon^{\boldsymbol{\beta}}_{ij}/z(\mathbf{W})$. Inserting these results into the stationarity condition (5.19) and combining with (5.33), we obtain

$$\left|\frac{2\omega_{ij}(\beta_i\beta_j-2\psi)-\beta_i\beta_j-2\psi\chi_{ij}}{z(\mathbf{W})}\right| \leq \frac{2\epsilon'_{ij}(\boldsymbol{\beta})+\epsilon_{ij}(\boldsymbol{\beta})}{z(\mathbf{W})}\,.$$

Dividing through by $2z(\mathbf{W})^{-1}\,|\beta_i\beta_j-2\psi|$ and rearranging yields

$$\left|\omega_{ij}-\frac{1}{2}\frac{f_{ij}(\boldsymbol{\beta})+\chi_{ij}}{1-f_{ij}(\boldsymbol{\beta})}\right| \leq \frac{2\epsilon'_{ij}(\boldsymbol{\beta})+\epsilon_{ij}(\boldsymbol{\beta})}{2\psi(1-f_{ij}(\boldsymbol{\beta}))}\,.$$

Finally, we apply Theorem 9 to approximate χ_{ij}, obtaining

$$\left|\omega_{ij} - \frac{f_{ij}(\boldsymbol{\beta})}{1 - f_{ij}(\boldsymbol{\beta})}\right| \leq \frac{2\epsilon'_{ij}(\boldsymbol{\beta}) + \epsilon_{ij}(\boldsymbol{\beta})}{2\psi(1 - f_{ij}(\boldsymbol{\beta}))} + \frac{1}{2}\frac{\epsilon_{ij}(\boldsymbol{\beta})}{1 - f_{ij}(\boldsymbol{\beta})}$$
$$= \frac{1}{1 - f_{ij}(\boldsymbol{\beta})}\left(\frac{2\epsilon'_{ij}(\boldsymbol{\beta}) + \epsilon_{ij}(\boldsymbol{\beta})}{2\psi} + \frac{\epsilon_{ij}(\boldsymbol{\beta})}{2}\right),$$

completing the proof.

5.7.2 Proof of Lemma 5

We will first derive the expression for the Jacobian of $\mathbf{h}$ given in the text as (5.26). For notational compactness, let $x_{ij} = f_{ij}(\boldsymbol{\beta})$. We first calculate

$$\frac{\partial x_{ij}}{\partial \beta_k} = \begin{cases} -\frac{x_{ij}}{2\psi} & k \neq i, j \\ x_{ij}(\frac{1}{\beta_i} - \frac{1}{2\psi}) & k = i \\ x_{ij}(\frac{1}{\beta_j} - \frac{1}{2\psi}) & k = j\,. \end{cases}$$

Next,

$$\frac{\partial}{\partial \beta_k}\left[\frac{x_{ij}}{1 - x_{ij}}\right] = \frac{1}{(1 - x_{ij})^2}\frac{\partial x_{ij}}{\partial \beta_k}\,.$$

We can therefore write the components of the Jacobian as

$$\frac{\partial h_i(\boldsymbol{\beta})}{\partial \beta_k} = \begin{cases} \left(\frac{1}{\beta_i} - \frac{1}{2\psi}\right)\sum_{j\neq i}\frac{x_{ij}}{(1-x_{ij})^2} & k = i \\ \frac{1}{\beta_k}\frac{f_{ik}}{(1-f_{ik})^2} - \frac{1}{2\psi}\sum_{j\neq i}\frac{x_{ij}}{(1-x_{ij})^2} & k \neq i\,. \end{cases}$$

It is convenient to define the matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ with entries $s_{ij} = \frac{x_{ij}}{(1-x_{ij})^2}$, in which case we can write

$$\frac{\partial h_i(\boldsymbol{\beta})}{\partial \beta_k} = \begin{cases} \left(\frac{1}{\beta_i} - \frac{1}{2\psi}\right)\sum_{j\neq i} s_{ij} & k = i \\ \frac{1}{\beta_k}a_{ik} - \frac{1}{2\psi}\sum_{j\neq i} s_{ij} & k \neq i\,. \end{cases}$$

Let $\mathbf{B} = \text{diag}\boldsymbol{\beta}$ and $\mathbf{D} = \text{diag}\mathbf{Ae}$. We can then write the Jacobian in matrix form as

$$\mathbf{J} = (\mathbf{S} + \mathbf{D})\mathbf{B}^{-1} - \frac{1}{2\psi}\mathbf{SE}\,.$$

We next note that $\mathbf{SE} = \mathbf{DE}$, and therefore substitute $\mathbf{SE} = \frac{1}{2}(\mathbf{SE} + \mathbf{DE})$. The result is

$$\mathbf{J} = (\mathbf{S} + \mathbf{D})\left(\mathbf{B}^{-1} - \frac{1}{4\psi}\mathbf{E}\right)\,,$$

as was to be shown.

We now proceed with the proof of Lemma 5. We will employ the crude lower bound (see [29]):

$$\lambda_1(\mathbf{J}) \geq \lambda_1\left(\mathbf{A} + \mathbf{D}\right)\lambda_1\left(\mathbf{B}^{-1} - \frac{1}{4\psi}\mathbf{E}\right)\,.$$

Starting with the second factor, we can obtain an analytical inverse (using, for example, the Sherman-Morrison formula):

$$\left(\mathbf{B}^{-1} - \frac{1}{4\psi}\mathbf{E}\right)^{-1} = \mathbf{B} + \frac{\mathbf{b}\mathbf{b}^T}{2\psi}\,.$$

We can separately upper bound the eigenvalues of each term on the righthand side, obtaining the bound

$$\lambda_n\left(\mathbf{B} + \frac{\mathbf{b}\mathbf{b}^T}{2\psi}\right) \leq \max_\ell \beta_\ell + \frac{1}{2\psi}\sum_\ell \beta_\ell^2 \leq n(n-1).$$

In the final inequality we have used that $\beta_n \leq n-1$ and $\boldsymbol{\beta} \geq \mathbf{e}$. Inverting this expression gives the first factor in the statement of the lemma.

Entrywise Taylor expansion, valid in $\mathcal{B}_\delta$, yields

$$\mathbf{A} + \mathbf{D} = (\mathbf{F} + \mathbf{G}_1) + 2(\mathbf{F}^{\circ 2} + \mathbf{G}_2) + 3(\mathbf{F}^{\circ 3} + \mathbf{G}_3) + \cdots \tag{5.34}$$

where $\mathbf{F} = \frac{\boldsymbol{\beta}\boldsymbol{\beta}^T}{2\psi}$, $\mathbf{F}^{\circ p}$ denotes Hadamard (entrywise) exponentiatian, and $\mathbf{G}_p$ is the

diagonal matrix with entries

$$[\mathbf{G}_p]_{ii} = \sum_{j=1}^{n} x_{ij}^p - 2x_{ii}^p \, .$$

Each term in the series (5.34) is weakly diagonally dominant, and therefore positive semi-definite by the Geršgorin disc theorem [105]. It will therefore suffice to bound below the eigenvalues of the first term. Since $\mathbf{F}$ is positive-semidefinite, we can obtain lower bounds in the eigenvalues by checking the eigenvalues of $\mathbf{G}_1$. Each entry is

$$[\mathbf{G}_1]_{ii} = \sum_{j=1}^{n} \frac{\beta_i \beta_j}{2\psi} - 2\frac{\beta_i^2}{2\psi} = \beta_i - 2\frac{\beta_i^2}{2\psi} = \beta_i \left(1 - \frac{\beta_i}{\psi}\right) > 0 \, .$$

We can bound this expression away from 0 by noting that

$$\beta_i \left(1 - \frac{\beta_i}{\psi}\right) \geq \min\left\{1 - \psi^{-1}, \beta_n \left(1 - \frac{\beta_n}{\psi}\right)\right\} \, . \tag{5.35}$$

By hypothesis, $2\psi \geq n \geq 5$, and therefore $\psi \geq 5/2$. The first argument of the minimum is thus no smaller than $\frac{3}{5}$. On the other hand, inserting the bound $\beta_n \leq \sqrt{2\psi - \delta} < \sqrt{2\psi}$, we can bound the second argument as

$$\beta_n \left(1 - \frac{\beta_n}{\psi}\right) \geq 1 - \frac{2}{\sqrt{5}} > 0 \, ,$$

Inserting this lower bound into (5.35) yields the result.

5.7.3 Proof of Theorem 13

It is convenient to define the function $c : [\beta_1, \beta_n] \to \mathbb{R}$ by

$$c_{\boldsymbol{\beta}}(z) = \sum_{\ell} \frac{z\beta_\ell}{2\psi - \beta_\ell z} - \frac{z^2}{2\psi - z^2} \, .$$

Note that $c_{\boldsymbol{\beta}}(\beta_i) = h_i(\boldsymbol{\beta})$. Then, supposing that $\boldsymbol{\beta}$ is a solution to (5.23), $c_{\boldsymbol{\beta}}(\beta_i) = d_i$. Additionally, if $\boldsymbol{\beta} \in \mathcal{B}_\delta$, then c is continuously differentiable on the interval $[\beta_1, \beta_n]$.

We will therefore show that $c'_{\boldsymbol{\beta}}(z) > 0$ for all z on this interval. The derivative is

$$c'_{\boldsymbol{\beta}}(z) = 2\psi \left(\sum_{\ell} \frac{\beta_\ell}{(2\psi - \beta_\ell z)^2} - \frac{z}{(2\psi - z^2)^2} \right) .$$

We will compute Taylor series of the terms inside the parentheses. Convergence of these series is guaranteed by the hypothesis that $\boldsymbol{\beta} \in \mathcal{B}_\delta$. It is convenient to introduce the notation $\psi_p = \frac{1}{2} \sum_\ell \beta_\ell^p$. Note that $\psi = \psi_1$.

First,

$$\sum_{\ell} \frac{\beta_\ell}{(2\psi - \beta_\ell z)^2} = \frac{1}{4\psi^2} \sum_{\ell=1}^{n} \beta_\ell \sum_{p=1}^{\infty} p \left(\frac{\beta_\ell z}{2\psi} \right)^{p-1} = \frac{1}{4\psi^2} \sum_{p=1}^{\infty} p \left(\frac{z}{2\psi} \right)^{p-1} (2\psi_p) .$$

Next,

$$\frac{z}{(2\psi - z^2)^2} = \frac{1}{4\psi^2} \sum_{p=1}^{\infty} p \left(\frac{z^2}{2\psi} \right)^{p-1} = \frac{1}{4\psi^2} \sum_{p=1}^{\infty} p \left(\frac{z}{2\psi} \right)^{p-1} z^{p-1} .$$

We thus write

$$c'_{\boldsymbol{\beta}}(z) = \frac{1}{2\psi} \sum_{p=1}^{\infty} p \left(\frac{z}{2\psi} \right)^{p-1} \left(2\psi_p - z^{p-1} \right) .$$

Since $z \le \beta_n$ by hypothesis, $z^{p-1} \le 2\psi_p$. Each term in the ezpansion is therefore strictly positive and we conclude that $c'_{\boldsymbol{\beta}}(z) > 0$. This proves (a).

To prove (b), we truncate to second order, obtaining

$$c'_{\boldsymbol{\beta}}(z) > \frac{1}{2\psi} \left(2\psi - 1 + \frac{z}{\psi} (2\psi_2 - z) \right) . \tag{5.36}$$

A small amount of algebra in combination with the hypotheses that $n \ge 5$ and $\mathbf{e} \le \boldsymbol{\beta} \le (n-1)\mathbf{e}$ shows that $\frac{z}{\psi}(2\psi_2 - z) \ge 1$ for $z \in [\beta_1, \beta_n]$. Inserting this lower bound into (5.36), we obtain

$$c'_{\boldsymbol{\beta}}(z) > 1 .$$

In particular, $c_{\boldsymbol{\beta}}^{-1}$ is a nonexpansive map. Since $c_{\boldsymbol{\beta}}^{-1} : d_i \mapsto \beta_i$, claim (b) follows.

To prove (c), it suffices to show that $\beta_1 \le d_1$ and apply (b). Since $\boldsymbol{\beta} \in \mathcal{B}_\delta$, $\beta_1 \ge 1$. We compute

$$d_1 = \beta_1 \sum_{\ell \neq 1} \frac{\beta_\ell}{2\psi - \beta_1 \beta_\ell} \ge \beta_1 \sum_{\ell \neq 1} \frac{\beta_\ell}{2\psi - \beta_1} = \beta_1 \ .$$

The first inequality uses $\beta_\ell \ge \beta_1$ and the hypothesis $\beta_1 \ge 1$.

Chapter 6

Looking Ahead

The research presented in the previous four chapters opens onto many interesting vistas of further development. I would like to highlight three: mechanistic network modeling; hypergraph methodology; and inference in diffusion processes on graphs.

Mechanistic Network Modeling

One promising direction of work concerns the role of generative models in network analysis. Several common tasks, such as community detection and core-periphery detection, can be formulated in terms of statistical generative models with tractable likelihood functions. These models often take the form of modified stochastic block-models. Tractable likelihoods enable practical statistical inference, allowing one to cast the results of analysis as principled solutions to statistically-grounded problems. On the other hand, the expressive power of these generative models are often constrained by independence assumptions. Mechanistic models of network formation, in which nodes and edges evolve in response to features of the current network state – can ameliorate this limitation while providing "physical" interpretations of quantities that would otherwise be merely statistical. This increase in expressive power often comes at a major cost: most such models are difficult to analyze and have intractable likelihoods. Approximations are required to understand their physical properties, and inference, if performed at all, must necessarily be approximate with several researcher

degrees of freedom.[1] The adaptive voter model discussed in Chapter 2 serves as an example of this trade-off. While the AVM joins a very small class of dynamical processes that display persistent community structure, the model possess an intractable likelihood function and can only be analyzed approximately in expectation.

However, scientists are increasingly collecting data sets describing networked systems evolving over time. In principle, these data sets can support mechanistic modeling and inference. An example of this kind of data is the flow of people – and therefore expertise and prestige – between academic institutions. Several recent papers have highlighted extreme prestige biases in faculty hiring and productivity, over and above that which can be explained by researcher merit [140, 210, 211, 54]. To our knowledge, however, there are no extant models of how hiring decisions are made in response to perceived prestige. In ongoing work, several collaborators and I are studying a simple, mechanistic model of agents endorsing other agents in response to universally-visible networked prestige scores. The mechanistic nature of the model offers a rich phenomenology, including a phase transition marking the emergence of prestige-based hierarchies. However, the model also possesses a tractable likelihood, allowing parameters to be inferred from data and compared between systems. An interesting feature of the model is that the evolving ranks are naturally interpreted as time-dependent centrality scores. Unlike in most other extant treatments of centrality in temporal networks [3, 199], these scores have concrete, mechanistic interpretation in terms of the dynamics of the system. The continuation of this research, as well as extensions of this approach to other analytical tasks such as community detection, appears to be an extremely rich area of future work. These and related prospects highlight the mutually-beneficial interaction of network theory and network data science.

[1] A roadmap for approximate Bayesian computation in dynamical processes on networks is laid out in [68].

Hypergraph Network Science

The development of hypergraph configuration models in Chapter 4 provides a promising launching-off point for further questions in the analysis of polyadic data. One of the most immediate opportunities for future work lies in mesoscale inference in polyadic models. There does exist a growing body of theory concerning mesoscale inference in hypergraphs, usually focusing on consistency, detectability, and recoverability results in planted-partition type models. Several of these papers develop tools indirectly, either by projecting the hypergraph [4] or by constructing a generative model for the associated bipartite graph [74, 121]. Still other analyses are restricted to uniform hypergraphs [46, 115, 114]; require known partition sizes [86, 85]; or have strong sparsity requirements [12] in order to guarantee their functioning. These results are of significant mathematical and statistical interest, but offer relatively little guidance for how an applied researcher should approach the problem of inferring mesoscopic structure in a given data set. A principled and highly general recipe for this task in the context of networks is offered by Bayesian stochastic blockmodel inference [163], which is usually performed by defining a Gibbs sampler over the posterior distribution of a stochastic blockmodel with suitably chosen priors. It would be of substantial applied interest to define an appropriate model class and implement a version of this solution. Another, related approach would be to define a mechanistic generative model, as discussed in the previous section, for hypergraph data sets. Such an approach would leverage the observation that many hypergraph data sets of practical interest consist of timestamped interactions [25]. The method of [27] for modeling sequences of sets may serve as a useful starting point for such a model.

A potentially-related question is that of hypergraph centrality. Many standard centrality scores take the form of eigenvectors of matrices associated with a given graph. This depends on the availability of a richly-developed spectral theory of the associated matrices. Hypergraphs, on the other hand, are often represented in the formalism of tensors. It is therefore necessary to either leverage the theory of tensors [24] or define dyadic projections [222] in order to apply spectral techniques. Both of

these approaches have substantial limitations – the first is only currently developed for uniform hypergraphs, while the second discards polyadic information entirely. More general approaches that gracefully incorporate polyadic interactions of varying sizes would constitute major advances in this field.

Learning Diffusion on Graphs

Diffusion on graphs is a well-studied problem, modeled in its most elementary form by the simple random walk. Many more complex definitions are possible; see [167] and accompanying citations for a review. One problem which to our knowledge has gone untreated is the *learning* of the parameters of a diffusion process from observations on a graph. While this problem is likely of fairly general interest, there is also opportunity to apply a candidate method to the data on spatial segregation used in Chapter 3. A diffusion process that best fit the available data would be of interest as a null model for analyzing the dynamics. Patterns in the data that significantly deviated from the predictions of the diffusion null would potentially be attributable to external forces, such as housing discrimination, discriminatory governmental policy, or geographic barriers, whose impact could then be analyzed and perhaps quantified. Such a project would pose considerable computational challenges. It is not obvious, for example, that a version of this model exists with a tractable likelihood. Further exploration on this topic is surely warranted.

Outlook

In the four chapters of this thesis, we have emphasized the opportunities that network science offers to applied mathematicians, especially the opportunities that have emerged from the field's gradual and ongoing movement toward data science. On the one hand, this trend has cast doubt on our ability to make universal statements about "Complex Networks" *tout court*, arguably decreasing the relevance of a certain kind of theory that emphasizes universality. On the other hand, the increasing availability

of networked data requires the development of tools with which to analyze it, as well as substantive theory of the data-generating processes. Especially exciting work can be expected at the intersection of these two bodies of theory, where knowledge of domain-relevant processes can inform subsequent analysis. Each of the extensions described above attempt to flesh out this intersection. There is plenty of work to do.

Bibliography

[1] 2014 American Community Survey 5-Year Estimates, Table B01003, 2016.

[2] Gregory Acs, Rolf Pendall, Mark Treskon, and Amy Khare. The Cost of Segregation. Technical report, The Urban Institute, 2017.

[3] Walid Ahmad, Mason A Porter, and Mariano Beguerisse-Díaz. Tie-decay temporal networks in continuous time and eigenvector-based centralities. *arXiv preprint arXiv:1805.00193*, 2018.

[4] Kwangjun Ahn, Kangwook Lee, and Changho Suh. Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):959–974, 2018.

[5] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed Membership Stochastic Blockmodels. Technical report, 2008.

[6] Georgios Amanatidis, Bradley Green, and Milena Mihail. Graphic realizations of joint-degree matrices. *arXiv preprint arXiv:1509.07076*, pages 1–18, 2015.

[7] Shun-Ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.

[8] Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2007.

[9] Antonio Ambrosetti and Paul H Rabinowitz. Dual variational methods in critical point theory and applications. *Journal of Functional Analysis*, 14(4):349–381, 1973.

[10] Aris Anagnostopoulos, Alessandro Bessi, Guido Caldarelli, Michela Del Vicario, Fabio Petroni, Antonio Scala, Fabiana Zollo, and Walter Quattrociocchi. Viral misinformation: The role of homophily and polarization. In *Proceedings of the 24th International Conference on World Wide Web*, pages 5–6, 2014.

[11] Omer Angel, Remco van der Hofstad, and Cecilia Holmgren. Limit laws for self-loops and multiple edges in the configuration model. *arXiv:1603.07172*, pages 1–19, 2016.

[12] Maria Chiara Angelini, Francesco Caltagirone, Florent Krzakala, and Lenka Zdeborová. Spectral detection on sparse hypergraphs. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 66–73. IEEE, 2015.

[13] Yael Artzy-Randrup and Lewi Stone. Generating uniformly distributed random networks. *Phys. Rev. E*, 72:056708, Nov 2005.

[14] K. B. Athreya and P. E. Ney. *Branching Processes*. Dover Books on Mathematics. Dover Publications, 2004.

[15] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, 2015.

[16] Albert-László Barabási and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):11, 1999.

[17] Albert-Laszlo Barabasi, Hawoong Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002.

[18] Riddhipratim Basu and Allan Sly. Evolving voter model on dense random graphs. *The Annals of Applied Probability*, 27(2):1235–1288, 2017.

[19] Michael Batty. Entropy and Spatial Geometry. *Royal Geographical Society*, 4(4):230–236, 1972.

[20] Michael Batty. Spatial Entropy. *Geographical Analysis*, 6(1):1–31, 1974.

[21] Michael Batty. Entropy in spatial aggregation. *Geographical Analysis*, 8(1):1–21, 1976.

[22] Michael Batty, Robin Morphet, Paolo Masucci, and Kiril Stanilov. Entropy, complexity, and spatial information. *Journal of Geographical Systems*, 16(4):363–385, 2014.

[23] Edward A Bender and E Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978.

[24] Austin R Benson. Three hypergraph eigenvector centralities. *SIAM Journal on Mathematics of Data Science*, 1(2):293–312, 2019.

[25] Austin R. Benson, Rediet Abebe, Michael T. Schaub, Ali Jadbabaie, and Jon Kleinberg. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*, 115(48):11221–11230, 2018.

[26] Austin R. Benson, David F. Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[27] Austin R Benson, Ravi Kumar, and Andrew Tomkins. Sequences of sets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1148–1157, 2018.

[28] Luis M. A. Bettencourt, Joe Hand, and José Lobo. Spatial Selection and the Statistics of Neighborhoods. 2015.

[29] Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013.

[30] Kshipra Bhawalkar, Sreenivas Gollapudi, and Kamesh Munagala. Coevolutionary opinion formation games. *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pages 41–50, 2013.

[31] James Bisgard. Mountain passes and saddle points. *SIAM Review*, 57(2):275–292, 2015.

[32] Joseph Blitzstein and Persi Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet mathematics*, 6(4):489–522, 2011.

[33] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal Of Statistical Mechanics-Theory And Experiment*, 10:1–12, 2008.

[34] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesús Gómez-Gardenes, Miguel Romance, Irene Sendina-Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[35] Gesa A. Böhme and Thilo Gross. Analytical calculation of fragmentation transitions in adaptive networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 83(3), 2011.

[36] Gesa A. Böhme and Thilo Gross. Fragmentation transitions in multistate voter models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 85(6), 2012.

[37] Béla Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.

[38] Jonathan R Bradley, Christopher K Wikle, and Scott H Holan. Regionalization of multiscale spatial processes using a criterion for spatial aggregation error. *Journal of the Royal Statistical Society: Series B (Statistial Methodology)*, 79(3):815–832, 2017.

[39] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On finding graph clusterings with maximum modularity. In *International Workshop on Graph-Theoretic Concepts in Computer Science*, pages 121–132. Springer, 2007.

[40] Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1 1967.

[41] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 1998.

[42] Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 10(1):1–10, 2019.

[43] Sonia Cafieri, Pierre Hansen, and Leo Liberti. Loops and multiple edges in modularity maximization of networks. *Physical Review E*, 81(4):046102, 2010.

[44] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[45] Corrie J Carstens. Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast curveball algorithm. *Physical Review E*, 91(4):042812, 2015.

[46] I Chien, Chung-Yi Lin, and I-Hsiang Wang. Community detection in hypergraphs: Optimal statistical limit and efficient algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 871–879, 2018.

[47] P. S. Chodrow. Structure and information in spatial segregation. *Proceedings of the National Academy of Sciences*, 114(44), 2017.

[48] P. S. Chodrow. Moments of uniformly random multigraphs with fixed degree sequences. *Submitted and under revision, arXiv: 1909.09037*, 2019.

[49] Philip Chodrow and Andrew Mellor. Annotated hypergraphs: models and applications. *Applied Network Science*, 5(1):9, 2020.

[50] Philip S. Chodrow and Peter J. Mucha. Markovian approximations for binary-state coevolving opinion networks. *Working Paper*, 2018.

[51] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

[52] Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145, 2002.

[53] William A V Clark, Eva Anderson, John Östh, and Bo Malmberg. A multiscalar analysis of neighborhood composition in Los Angeles, 2000–2010: A location-based approach to segregation and diversity. *Annals of the Association of American Geographers*, 105(6):1260–1284, 2015.

[54] Aaron Clauset, Samuel Arbesman, and Daniel B Larremore. Systematic inequality and hierarchy in faculty hiring networks. *Science advances*, 1(1):e1400005, 2015.

[55] Aaron Clauset, C R Shalizi, and Mark E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[56] Peter Clifford and Aidan Sudbury. A model for spatial conflict. *Biometrika*, 60(3):581–588, 1973.

[57] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nature Physics*, 2(2):110–115, 2006.

[58] Owen T. Courtney and Ginestra Bianconi. Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes. *Physical Review E*, 93(6):1–26, 2016.

[59] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[60] Imre Csiszar and Paul C Shields. Information Theory and Statistics: A Tutorial. *Foundations and Trends™ in Communications and Information Theory*, 1(4):417–528, 2004.

[61] Charo I Del Genio, Hyunju Kim, Zoltán Toroczkai, and Kevin E Bassler. Efficient and exact sampling of simple graphs with given arbitrary degree sequence. *PloS one*, 5(4):e10012, 2010.

[62] J. C. Delvenne, S. N. Yaliraki, and M. Barahona. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, 2010.

[63] G. Demirel, F. Vazquez, Gesa A. Böhme, and T. Gross. Moment-closure approximations for discrete adaptive networks. *Physica D: Nonlinear Phenomena*, 267:68–80, 2014.

[64] Güven Demirel, Edmund Barter, and Thilo Gross. Dynamics of epidemic diseases on a growing adaptive network. *Scientific Reports*, 7:42352, Feb 2017.

[65] Inderjit S Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 2003.

[66] Juan C Duque and Sergio J Rey. The max-p regions problem. *Journal of Regional Science*, 52(3):397–419, 2012.

[67] Richard Durrett, James P. Gleeson, Alun L. Lloyd, Peter J. Mucha, Feng Shi, David Sivakoff, Joshua E. S. Socolar, and Chris Varghese. Graph fission in an evolving voter model. *Proceedings of the National Academy of Sciences of the United States of America*, 109(10):3682–7, 2012.

[68] Ritabrata Dutta, Antonietta Mira, and Jukka-Pekka Onnela. Bayesian inference of spreading processes on networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2215):20180129, 2018.

[69] P Erdos and A Renyi. On Random Graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[70] Péter L Erdős, Catherine Greenhill, Tamás Róbert Mezei, István Miklós, Dániel Soltész, and Lajos Soukup. The mixing time of the swap (switch) markov chains: a unified approach. *arXiv:1903.06600*, 2019.

[71] Paul Expert, Tim S Evans, Vincent D Blondel, and Renaud Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19):7663–7668, 2011.

[72] Glenn Firebaugh and Francesco Acciai. For Blacks in America, the Gap in Neighborhood Poverty has Declined Faster Than Segregation. *Proceedings of the National Academy of Sciences*, page 201607220, 2016.

[73] Glenn Firebaugh and Chad R. Farrell. Still large, but narrowing: The sizable decline in racial neighborhood inequality in metropolitan America, 1980–2010. *Demography*, 53(1):139–164, 2016.

[74] Laura Florescu and Will Perkins. Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pages 943–959, 2016.

[75] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2006.

[76] Bailey K Fosdick, Daniel B Larremore, Joel Nishimura, and Johan Ugander. Configuring random graph models with fixed degree sequences. *SIAM Review*, 60(2):315–355, 2018.

[77] Christopher S. Fowler, Barrett A. Lee, and Stephen A. Matthews. The contributions of places to metropolitan ethnoracial diversity and segregation: Decomposing change across space and time. *Demography*, 53(6):1955–1977, 2016.

[78] James H. Fowler. Connecting the Congress: A study of cosponsorship networks. *Political Analysis*, 14(04):456–487, 2006.

[79] James H Fowler. Legislative cosponsorship networks in the U.S. House and Senate. *Social Networks*, 28:454–465, 2006.

[80] David M. Frankel and Oscar Volij. Measuring school segregation. *Journal of Economic Theory*, 146(1):1–38, 2011.

[81] William H. Frey and Dowell Myers. Racial Segregation in US Metropolitan Areas and Cities, 1990–2000: Patterns, Trends, and Explanations. (April):1–66, 2005.

[82] Matias Garreton and Raimundo Sánchez. Identifying an optimal analysis level in multiscalar regionalization: A study case of social distress in Greater Santiago. *Computers, Environment and Urban Systems*, 56:14–24, 2016.

[83] Martin Gerlach and Eduardo G Altmann. Testing statistical laws in complex systems. *Physical review letters*, 122(16):168301, 2019.

[84] Gourab Ghoshal, Vinko Zlatić, Guido Caldarelli, and Mark E. J. Newman. Random hypergraphs and their applications. *Physical Review E*, 2009.

[85] Debarghya Ghoshdastidar and Ambedkar Dukkipati. Consistency of spectral partitioning of uniform hypergraphs under planted partition model. In *Advances in Neural Information Processing Systems*, pages 397–405, 2014.

[86] Debarghya Ghoshdastidar, Ambedkar Dukkipati, et al. Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics*, 45(1):289–315, 2017.

[87] Chad Giusti, Robert Ghrist, and Danielle S. Bassett. Two's company, three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data. *Journal of Computational Neuroscience*, 41(1):1–14, 2016.

[88] James P. Gleeson. High-accuracy approximation of binary-state dynamics on networks. *Physical Review Letters*, 107(6):1–9, 2011.

[89] Ezra Haber Glenn. acs: Download, manipulate, and present American Community Survey and Decennial data from the US Census, 2016.

[90] Benjamin H Good, Yves-Alexandre De Montjoye, and Aaron Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010.

[91] Boris L. Granovsky and Neal Madras. The noisy voter model. *Stochastic Processes And Their Applications*, 55(1):23–43, 1995.

[92] Catherine Greenhill. A polynomial bound on the mixing time of a markov chain for sampling regular directed graphs. *The Electronic Journal of Combinatorics*, 18(1):234, 2011.

[93] Catherine Greenhill. The switch markov chain for sampling irregular graphs. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1564–1572. SIAM, 2014.

[94] Jacopo Grilli, György Barabás, Matthew J. Michalska-Smith, and Stefano Allesina. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, pages 210–213, 2017.

[95] Thilo Gross and Bernd Blasius. Adaptive coevolutionary networks: A review. *Journal of The Royal Society Interface*, 5(20):259–271, Mar 2008.

[96] Thilo Gross, Carlos J. Dommar D"Lima, and Bernd Blasius. Epidemic dynamics on an adaptive network. *Physical Review Letters*, 96(20):208701, May 2006.

[97] Thilo Gross and Hiroki Sayama. *Adaptive Networks: Theory, Models and Applications*. Springer Science & Business Media, 2009.

[98] Pontus Hennerdal and Michael Meinild Nielsen. A multiscalar approach for identifying clusters and segregation patterns that avoids the Modifiable Areal Unit Problem. *Annals of the American Association of Geographers*, 4452(June):1–20, 2017.

[99] A. D. Henry, P. Pralat, and C.-Q. Zhang. Emergence of segregation in evolving social networks. *Proceedings of the National Academy of Sciences*, 108(21):8605–8610, 2011.

[100] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[101] Richard A. Holley and Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *The Annals of Probability*, 3(4):643–663, 1975.

[102] Petter Holme. Rare and everywhere: Perspectives on scale-free networks. *Nature Communications*, 10(1):1–3, 2019.

[103] Petter Holme and M. E. J. Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 74(5):1–5, 2006.

[104] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[105] Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

[106] Leonhard Horstmeyer, Christian Kuehn, and Stefan Thurner. Network topology near criticality in adaptive epidemics. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 98(4):042313, 2018.

[107] Youssef Jabri. *The Mountain Pass Theorem: Variants, Generalizations and Some Applications*, volume 95. Cambridge University Press, 2003.

[108] Paul A Jargowsky and Jeongdai Kim. A measure of spatial segregation: The generalized neighborhood sorting index. In *National Poverty Center Working Paper Series*, number 05-3. 2005.

[109] Mark Jerrum and Alistair Sinclair. Fast uniform generation of regular graphs. *Theoretical Computer Science*, 73(1):91–100, 1990.

[110] M. Ji, C. Xu, C. W. Choi, and P. M. Hui. Correlations and analytical approaches to co-evolving voter models. *New Journal of Physics*, 15, 2013.

[111] Bogumil Kaminski, Valerie Poulin, Pawel Pralat, Przemyslaw Szufel, and Francois Theberge. Clustering via hypergraph modularity. *arXiv:1810.04816*, pages 1–17, 2018.

[112] Ravi Kannan, Prasad Tetali, and Santosh Vempala. Simple markov-chain algorithms for generating bipartite graphs and tournaments. *Random Structures & Algorithms*, 14(4):293–308, 1999.

[113] Brian Karrer and Mark E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 83(1):1–11, 2011.

[114] Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Community detection in hypergraphs, spiked tensor models, and sum-of-squares. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 124–128. IEEE, 2017.

[115] Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*, 2018.

[116] Daichi Kimura and Yoshinori Hayakawa. Coevolutionary networks with homophily and heterophily. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 78(1):1–7, 2008.

[117] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

[118] Bryan Klimt and Yiming Yang. Introducing the Enron Corpus. In *CEAS*, 2004.

[119] Tarun Kumar, Sankaran Vaidyanathan, Harini Ananthapadmanabhan, Srinivasan Parthasarathy, and Balaraman Ravindran. Hypergraph clustering: a modularity maximization approach. *arXiv:1812.10869*, 2018.

[120] Yacoub H. Kureh and Mason A. Porter. Fitting in and breaking up: A nonlinear version of coevolving voter models. *arXiv:1907.11608*, 2019.

[121] Daniel B Larremore, Aaron Clauset, and Abigail Z Jacobs. Efficiently inferring community structure in bipartite networks. *Physical Review E*, 90(1):012805, 2014.

[122] Barret A Lee, Sean F. Reardon, Glenn Firebaugh, Chad R. Farrell, Stephen A. Matthews, and David O'Sullivan. Beyond the census tract: Patterns and determinants of racial segregation at multiple geographic scales. *American Sociological Review*, 73(5):766–791, 2008.

[123] Hsuan-Wei Lee, Nishant Malik, and Peter J. Mucha. Evolutionary prisoner's dilemma games coevolving on adaptive networks. *Journal of Complex Networks*, 6(1):1–23, Feb 2018.

[124] Hsuan-Wei Lee, Nishant Malik, Feng Shi, and Peter J Mucha. Social clustering in epidemic spread on coevolving networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 99(6):062301, 2019.

[125] Daniel T Lichter, Domenico Parisi, and Michael C Taquino. Toward a new macro-segregation? Decomposing segregation within and between metropolitan cities and suburbs. *American Sociological Review*, 80(4):843–873, 2015.

[126] May Lim, Richard Metzler, and Yaneer Bar-Yam. Global pattern formation and ethnic/cultural violence. *Science*, 317(5844):1540–1544, 2007.

[127] John R. Logan, Brian J. Stults, and Reynolds Farley. Segregation of minorities in the metropolis: Two decades of change. *Demography*, 41(1):1–22, 2004.

[128] John R. Logan, Zengwang Xu, and Brian J. Stults. Interpolating U.S. Decennial Census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer*, 66(3):412–420, 2014.

[129] Rémi Louf and Marc Barthélemy. Patterns of residential segregation. *arXiv.org*, pages 1–17, 2015.

[130] Ulrike Von Luxburg. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(March):395–416, 2007.

[131] Nishant Malik, Feng Shi, Hsuan Wei Lee, and Peter J. Mucha. Transitivity reinforcement in the coevolving voter model. *Chaos*, 26(12), 2016.

[132] Vincent Marceau, Pierre André Noël, Laurent Hébert-Dufresne, Antoine Allard, and Louis J. Dubé. Adaptive networks: coevolution of disease and topology. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 82(3), 2010.

[133] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE*, 10(9):1–26, 2015.

[134] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLOS ONE*, 10(9), 2015.

[135] Brendan D McKay and Nicholas C Wormald. Uniform generation of random regular graphs of moderate degree. *Journal of Algorithms*, 11(1):52–67, 1990.

[136] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[137] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.

[138] Michael Molloy and Bruce Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability, and Computing*, 7(3):295–305, 1998.

[139] Ray Monk. *Ludwig Wittgenstein: The Duty of Genius*. Random House, 2012.

[140] Allison C Morgan, Dimitrios J Economou, Samuel F Way, and Aaron Clauset. Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science*, 7(1):40, 2018.

[141] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[142] James Munkres. *Analysis on Manifolds*. Westview Press, 1997.

[143] M E J Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[144] M. E. J. Newman. Community detection in networks: Modularity optimization and maximum likelihood are equivalent. *arXiv.org*, pages 1–8, 2016.

[145] M. E. J. Newman and Aaron Clauset. Structure and inference in annotated networks. *Nature Communications*, 7(May):1–16, 2016.

[146] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004.

[147] Mark E. J. Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1):1–8, 2001.

[148] Mark E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):1–5, 2002.

[149] Mark E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):13, 2003.

[150] Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[151] Mark E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):17, 2001.

[152] Mark EJ Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[153] Joel Nishimura. The connectivity of graphs of graphs with self-loops and a given degree sequence. *Journal of Complex Networks*, 6(6):927–947, 2018.

[154] Richard Nock, Pascal Vaillant, Claudia Henry, and Frank Nielsen. Soft memberships for spectral clustering, with application to permeable language distinction. 42:43–53, 2009.

[155] S. Openshaw and P.J. Taylor. The modifiable areal unit problem. In N. Wrigley and R.J. Bennett, editors, *Quantitative Geography: A British View*. Routledge and Kegan Paul, London, 1981.

[156] J. Östh, W. A. V. Clark, and B. Malmberg. Measuring the scale of segregation using k-nearest neighbor aggregates. *Geographical Analysis*, 47:34–49, 2014.

[157] J. Östh, B. Malmberg, and E. Andersson. Analysing segregation using individualized neighborhoods. In C. D. Lloyd, I.G. Shuttleworth, and David W. S. Wong, editors, *Socio-spatial segregation: Concepts, processes, and outcomes*, pages 135–162. The Policy Press, Bristol, UK, 2015.

[158] David O'Sullivan and David W. S. Wong. A surface-based approach to measuring spatial segregation. *Geographical Analysis*, 39(2):147–168, 2007.

[159] Jorge M Pacheco, Arne Traulsen, and Martin A Nowak. Coevolution of strategy and structure in complex networks with dynamical linking. *Physical Review Letters*, 97(25):258103, 2006.

[160] Alice Patania, Giovanni Petri, and Francesco Vaccarino. The shape of collaborations. *EPJ Data Science*, 6(1):1–16, 2017.

[161] Leto Peel, Daniel B Larremore, and Aaron Clauset. The ground truth about metadata and community detection in networks. *Science Advances*, 3(5):e1602548, 2017.

[162] Tiago P. Peixoto. Nonparametric weighted stochastic block models. pages 1–20, 2017.

[163] Tiago P Peixoto. Bayesian stochastic blockmodeling. *Advances in Network Clustering and Blockmodeling*, pages 289–332, 2019.

[164] Pew Research Center. Modern Immigration Wave Brings 59 Million to U.S., Driving Population Growth and Change Through 2065: Views of Immigration's Impact on U.S. Society Mixed. pages 1–127, 2015.

[165] Flávio L Pinheiro, Francisco C Santos, and Jorge M Pacheco. Linking individual and collective behavior in adaptive social networks. *Physical Review Letters*, 116(12):128702, 2016.

[166] Mason A Porter. Nonlinearity+ networks: A 2020 vision. *arXiv preprint arXiv:1911.03805*, 2019.

[167] Mason A Porter and James P Gleeson. Dynamical systems on networks. 2016.

[168] Mason A Porter, Peter J Mucha, M E J Newman, and Casey M Warmbrand. A network analysis of committees in the U. S. House of Representatives. *Proceedings of the National Academy of Sciences*, 102(20):7057–7062, 2005.

[169] Sean F. Reardon. Measures of ordinal segregation. In Yves Flückiger, Sean F. Reardon, and Jacques Silber, editors, *Occupational and Residential Segregation (Research on Economic Inequality, Volume 17)*, pages 129 – 155. 2008.

[170] Sean F. Reardon, Chad R. Farrell, Stephen A. Matthews, David O'Sullivan, Kendra Bischoff, and Glenn Firebaugh. Race and space in the 1990s: Changes in the geographic scale of racial residential segregation, 1990-2000. *Social Science Research*, 38(1):55–70, 2009.

[171] Sean F. Reardon and G. Firebaugh. Measures of multigroup segregation. *Sociological Methodology*, 32:33–67, 2002.

[172] Sean F Reardon, Stephen A Matthews, David O Sullivan, and Barrett A Lee. The geographic scale of metropolitan racial segregation. 45(3):489–514, 2008.

[173] Sean F. Reardon and David O'Sullivan. Measures of spatial segregation. *Sociological Methodology*, 34(1):121–162, 2004.

[174] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, 2006.

[175] Elizabeth Roberto. Measuring inequality and segregation arxiv:1508.01167. pages 1–26, 2015.

[176] Elizabeth Roberto. The Spatial Context of Residential Segregation. *arXiv.org*, pages 1–27, 2016.

[177] Tim Rogers and Thilo Gross. Consensus time and conformity in the adaptive voter model. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 88(3), 2013.

[178] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[179] Hernán D. Rozenfeld, Diego Rybski, José S. Andrade, Michael Batty, H. Eugene Stanley, and Hernán a Makse. Laws of population growth. *Proceedings of the National Academy of Sciences*, 105(48):18702–18707, 2008.

[180] Hernán D. Rozenfeld, Diego Rybski, Xavier Gabaix, and Hernán a Makse. The area and population of cities: New insights from a different perspective on cities. *American Economic Review*, 101(5):2205–2225, 2011.

[181] Christopher S. Fowler. Segregation as a multiscalar phenomenon and its implications for neighborhood-scale research: the case of South Seattle 1990–2010. *Urban Geography*, 37(1):1–25, 2016.

[182] Fabio Saracco, Riccardo Di Clemente, Andrea Gabrielli, and Tiziano Squartini. Randomizing bipartite networks: the case of the World Trade Web. *Scientific Reports*, 5(10595):1–18, 2015.

[183] Michael T Schaub, Austin R Benson, Paul Horn, Gabor Lippner, and Ali Jadbabaie. Random walks on simplicial complexes and the normalized Hodge Laplacian. *arXiv:1807.05044*, pages 1–36, 2018.

[184] Thomas C. Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1(May):143–186, 1979.

[185] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[186] Feng Shi, Peter J. Mucha, and Richard Durrett. Multiopinion coevolving voter model with infinitely many phase transitions. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 88(6):1–15, 2013.

[187] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[188] Jianbo Shi and Jitendra Malik. Normalized Cuts and Image Segmentation. 22(8):888–905, 2000.

[189] Holly Silk, Güven Demirel, Martin Homer, and Thilo Gross. Exploring the adaptive voter model dynamics with a mathematical triple jump. *New Journal of Physics*, 16, 2014.

[190] Filippo Simini, Marta C. González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[191] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web*. ACM Press, 2015.

[192] Jonas Sjöstrand. Making multigraphs simple by a sequence of double edge swaps. *arXiv:1904.06999*, (April):1–11, 2019.

[193] Seth E Spielman and John R. Logan. Using high-resolution population data to identify neighborhoods and establish their boundaries. *Annals of the Association of American Geographers*, 103(1):67–84, 2013.

[194] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, and Philippe Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):e23176, 2011.

[195] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean François Pinton, Marco Quaggiotto, Wouter van den Broeck, Corinne Régis, Bruno Lina, and Philippe Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):1–13, 2011.

[196] S. H. Strogatz and D. J. Watts. Collective dynamics of "small-world" networks. *Nature*, 393(June):440–442, 1998.

[197] Giovanni Strona, Domenico Nappo, Francesco Boccacci, Simone Fattorini, and Jesus San-Miguel-Ayanz. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature communications*, 5:4114, 2014.

[198] Michael PH Stumpf and Mason A Porter. Critical truths about power laws. *Science*, 335(6069):665–666, 2012.

[199] Dane Taylor, Sean A Myers, Aaron Clauset, Mason A Porter, and Peter J Mucha. Eigenvector-based centrality measures for temporal networks. *Multiscale Modeling & Simulation*, 15(1):537–574, 2017.

[200] Henri Theil and Anthony J. Finezza. A note on the measurement of racial integration of schools by means of informational concepts. *Journal of Mathematical Sociology*, 1:187–194, 1971.

[201] J Toruniewska, K. Kułakowski, K Suchecki, and J. A. Hołyst. Coupling of link- and node-ordering in the coevolving voter model. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 96(4), 2017.

[202] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, 2012.

[203] Federico Vazquez and Víctor M Eguíluz. Analytical solution of the voter model on uncorrelated networks. *New Journal of Physics*, 10(6):1–19, 2008.

[204] Federico Vazquez, Víctor M Eguíluz, and Maxi San Miguel. Generic absorbing transition in coevolution dynamics. *Physical Review Letters*, 100(10), 2008.

[205] Norman D Verhelst. An efficient mcmc algorithm to sample binary matrices with fixed marginals. *Psychometrika*, 73(4):705, 2008.

[206] Fabien Viger and Matthieu Latapy. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. In *International Computing and Combinatorics Conference*, pages 440–449. Springer, 2005.

[207] Ivan Voitalov, Pim van der Hoorn, Remco van der Hofstad, and Dmitri Krioukov. Scale-free networks well done. *Physical Review Research*, 1(3):033034, 2019.

[208] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[209] Kyle Walker. tigris: Load Census TIGER/line shapefiles into R, 2016.

[210] Samuel F Way, Daniel B Larremore, and Aaron Clauset. Gender, productivity, and prestige in computer science faculty hiring networks. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1169–1179, 2016.

[211] Samuel F Way, Allison C Morgan, Daniel B Larremore, and Aaron Clauset. Productivity, prominence, and the effects of academic environment. *Proceedings of the National Academy of Sciences*, 116(22):10729–10733, 2019.

[212] Michael J. Webber. *Information Theory and Urban Spatial Structure*. Croon Helm, London, 1979.

[213] David W. S. Wong. Geostatistics As Measures of Spatial Segregation. *Urban Geography*, 20(7):635–647, 1999.

[214] David W. S. Wong. Comparing Traditional and Spatial Segregation Measures: A Spatial Scale Perspective. *Urban Geography*, 25(1):66–82, 2004.

[215] Su Do Yi, Seung Ki Baek, Chen Ping Zhu, and Beom Jun Kim. Phase transition in a coevolving network of conformist and contrarian voters. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 87(1):1–11, 2013.

[216] Hao Yin, Austin R. Benson, and Jure Leskovec. Higher-order clustering in networks. *Physical Review E*, 97(5), 2018.

[217] Hyejin Youn, Deborah Strumsky, Luis M. A. Bettencourt, and José Lobo. Invention as a combinatorial process: evidence from US patents. *Journal of the Royal Society, Interface*, 12(106):1–8, 2015.

[218] Jean Gabriel Young, Giovanni Petri, Francesco Vaccarino, and Alice Patania. Construction of and efficient sampling from the simplicial configuration model. *Physical Review E*, 96(3):1–6, 2017.

[219] Shuai Yuan, Pang Ning Tan, Kendra Spence Cheruvelil, Sarah M. Collins, and Patricia A. Soranno. Constrained spectral clustering for regionalization: Exploring the trade-off between spatial contiguity and landscape homogeneity. *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*, 2015.

[220] Pan Zhang and Cristopher Moore. Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *Proceedings of the National Academy of Sciences*, 111(51):18144–18149, 2014.

[221] Xiao Zhang and M. E. J. Newman. Multiway spectral community detection in networks. *Phys. Rev. E*, 92:052808, Nov 2015.

[222] Dengyong Zhou and Jiayuan Huang. Learning with hypergraphs: Clustering, classification, and embedding. In *NIPS*, 2007.

[223] Gerd Zschaler, Gesa A. Böhme, Michael Seißinger, Cristián Huepe, and Thilo Gross. Early fragmentation in the adaptive voter model on directed networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 85(4):046107, Apr 2012.