Studies in the Analysis of Stochastic Processes

by

Peter E. Wills

B.A., Reed College, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Applied Mathematics

2018

This thesis entitled:
Studies in the Analysis of Stochastic Processes
written by Peter E. Wills
has been approved for the Department of Applied Mathematics

_____

Prof. François Meyer

_____

Prof. Mark Hoefer

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Wills, Peter E. (Ph.D., Applied Mathematics)

Studies in the Analysis of Stochastic Processes

Thesis directed by Prof. François Meyer

As the quantity of recorded data grows exponentially, the development of techniques for analysis of such data have become a popular topic throughout industrial and academic research. However, the practical application of modern analytical methods must eventually confront the challenge of noisy data. The successful application of such methods is predicated on the ability to understand, quantiy, and anticipate the effects of random fluctuations in the inputs. This thesis is a compilation of three distinct projects, which are connected by the theme of understanding stochastic processes and extracting their essential information.

In the first project, we analyze the efficacy of an approach for hypothesis testing and confidence interval construction, which can be seen as an application of test supermartingales [130]. This approach is flexible, and in particular can be applied in tandem with arbitrary stopping rules, so that the number of trials to be performed need not be fixed in advance of the experiment. The method was originally developed for application in tests of local realism, which is a foundational concept in quantum mechanics [162, 163].

The present work examines the efficacy of the method when applied to parameter estimation based on repeated samples of a Bernoulli random variable. The relative simplicity of this scenario allows for benchmarking against an optimal technique (based on exact calculation of tail probabilities) and for rigorous analysis of the inversion problem necessary to generate confidence intervals based on our hypothesis tests. We show that our test supermartingale method attains an asymptotically optimal gain rate, which is the exponential rate of decay of the resulting p-value as the number of trials increases. We also show that the separation between the endpoint of a one-sided confidence interval and the true probability of success is $\mathcal{O}(\sqrt{\log(n)/n})$, while the optimal endpoint separation is $\mathcal{O}(1/\sqrt{n})$. This $\mathcal{O}(\sqrt{\log n})$ difference can be viewed as the cost of

the robustness against stopping rules. We propose an adaptive modification to our method which yields an $\mathcal{O}(1/\sqrt{n})$ endpoint separation when the number of trials is known in advance. This work resulted in the publication [156].

The second project examines the effect of thermal perturbations on standing wave structures that arise in thin-film magnetics. These structures, known as magnetic droplet solitons, or simply droplets, have been observed experimentally [9, 106] and thoroughly studied through the lens of partial differential equations [24, 36, 98]. Our analysis extends the analysis begun in [24], using the machinery of stochastic partial differential equations to quantify the effects of physical sources of randomness on the linewidth of the observed droplet. We obtain an analytical expression for the linewidth based on the linearized equations of motion, and use numerical simulations to compare this to the linewidth generated by the full nonlinear model. Along the way, we uncover a deterministic regime of drift instability missed by previous analyses [23, 24]. This work resulted in the publication [155]

The third project examines methods for pairwise comparison of graphs, focusing in particular on a method based on the effective graph resistance [48], referred to as the resistance-perturbation distance, or simply the resistance distance. Previous work [107] has established basic properties of the resistance distance; we continue that program by extending the method to be applicable to graphs of different sizes, and examining the efficacy of the method in detecting transitions in the community structure of a dynamic random graph model. We show that in order for the resistance distance to effective discern these transitions in community structure, the number of cross-community edges must be asymptotically dominated by the mean degree of the graph. In order to do this, we establish an asymptotic expression for the graph resistance between vertices in the stochastic blockmodel. This work resulted in the publication [158].

Finally, we perform a thorough study of the performance of various common graph comparison tools when used to detect common global and local structural discrepancies in graphs. These structures are typified by various random graph models, such as the stochatic blockmodel [2], preferential attachment [8], and Watts-Strogatz [153] models. Our results indicate that spectral

tools offer the most performant option when comparing graphs over multiple structural scales. Other graph distances are either too myopic (the graph edit distance, the resistance perturbation distance) or miss essential structure at the local level (comparison of principal eigenvalues). Overall, full comparison of the adjacency spectrum proves to be the most robust and universally applicable graph comparison tool across these experiments.

We also look at the performance of these graph comparison tools on two empirical datasets; a social contact network generated by students at a French primary school [138], and the functional connectome of the human brain, measured in subjects with and without autism spectrum disorder. We find that while spectral tools such as the adjacency spectral distance shows the best performance in population comparison experiments, vertex-comparison methods such as the resistance distance and DeltaCon [82] show superior performance when applied to anomaly detection in dynamic graphs.

Dedication

For Kathleen, whose love keeps me warm. For my mother, who gave me the foundation. For Momo, who soothes me when nothing else can.

Acknowledgements

First and foremost, I'd like to thank my advisors, Prof. François Meyer, Prof. Mark Hoefer, and Dr. Emanuel Knill. Their inexhaustible patience and insightful guidance has been the backbone of my success in the program, and I'm luck to have had the chance to work with and learn from them.

I'd also like to thank the supportive community in the Applied Mathematics department, who have provided me with financial support as a teaching assistant, and personal support and guidance throughout my years at CU. I'd like to thank in particular my committee members: Prof. Juan Restrepo, Prof. Manuel Lladser, and Prof. Sean O'Rourke of the Mathematics department.

Finally, I want to thank my friends and family who provided me with the support that got me through the difficult times. To Ahmed, Jay, Drew, Jason, and my sister Emily: thank you.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

The proliferation of cost-efficient data storage has fueled growing interest in analytical techniques applicable to large data sets. However, the effects of randomness within these techniques is often ignored [101]. In this thesis, we examine three distinct application areas, which are connected via the theme of understanding the behavior of stochastic processes.

Chapter 2 is focused on furthering the analytical understanding of a statistical tool originally developed to analyze the output of Bell tests, which are experimental tests of local realism, a fundamental theoretical aspect of quantum mechanics. One example of a recent Bell test experiment can be found in [132]. The foundational nature of the property being tested means that the experiments require extremely low p-values in order to unambiguously exclude alternative explanations [43].

Bell tests consist of repeated samples of some physical quantity B which will be negative under the null hypothesis that local realism holds, and positive if local realism is violated, as quantum theory predicts. The usual analytical method is to look at the sample mean $\hat{B}$ and sample deviation $\hat{\sigma}$, with $\hat{B} \gg \hat{\sigma}$ constituting strong evidence against the null hypothesis that $B < 0$ [89, 162]. However, this method suffers from problems. First, the trials are observably not independent and identically distributed. Second, the convergence in distribution of the central limit theorem, which justifies Gaussian approximations in general, does not hold in this case, as the set whose measure we are estimating is not fixed; in the case where the null hypothesis is false, as we gather more data, we move further and further into the extreme tail of the distribution proposed by the null hypothesis.

In [162], a method was developed that rigorously establishes p-values for Bell tests. These p-values make no assumptions that the trials are i.i.d., and do not rely on Gaussian approximations of the null distribution. The method can be seen as an application of the test supermartingales introduced in [130]. In short, if we construct a supermartingale V_n based on the first n trials, it's expectation is bounded above by 1, and so its inverse $p = 1/V_n$ is a p-value bound via Markov's inequality. We construct a particular test supermartingale, for which the resulting p-values can achieve an asymptotically optimal gain rate, which is to the rate of exponential decay of p_n as n increases. Recently, this technique was used in a Bell test scenario in which other methods proved unsuccessful [33].

Our work is an attempt to understand the behavior of this test-supermartingale method in a scenario that is analytically tractable. We focus on parameter estimation with repeated sampling of a Bernoulli random variable. In such a simplified situation, we are able to compare test supermartingale-based p-value bounds against the standard large-deviation bound [31, 69] and the exact p-value generated from the cumulative distribution. Furthermore, we can apply an approximate inversion technique in order to understand the confidence intervals generated by each method. This work led to [156].

In Chapter 3 we examine the influence of thermal noise on a particular localized structure arising in thin magnetic films. Magnetic nanostructures have potential applications in logic and storage [112] and in communications [165]. We focus on magnetic droplet silitons, or simply droplets. These structures exist in magnetic thin films for which the easy axis lies perpendicular to the plane of the film, which balances the exchange energy and allows for the development of stable localized structures [81].

Droplets have been observed experimentally in various scenarios [105, 9, 92, 36] and predicted theoretically using both full and asymptotically reduced models [100, 24, 67, 68]. The nucleation of droplets is driven by a nanocontact spin-torque oscillator, and theoretical analysis predicts that the magnetic field sharply transitions at the boundary of the oscillator. However, direct imaging of droplets has failed to produce the sharp transitions that are predicted theoretically [105].

We continue the theoretical analysis of droplets by introducing thermal noise into the dynamical equations governing droplet evoluion. In doing so, we hope to address the influence of thermal noise on both linewidth and droplet center variance, thus addressing the diffuse structural observations of [105]. The application of soliton perturbation theory to this system of stochastic partial differential equations allows for the analytical derivation of both the center variance and the generation linewidth [134]. We pair this analysis with numerical simulation of both the reduced and the full models. We also analyze the deterministic stability of the droplet, uncovering further potential sources of imaging noise. This work let to the publication of [155].

In Chapter 4 we analyze the performance of the effective graph resistance [48] in detecting topological changes in the dynamic stochastic blockmodel. The study of time-varying graphs is fundamental to myriad disciplines, including computer network security [70], social network analysis [5], neuroscience [126], and biology [15], with the detection of anomalous changes in the evolution of dynamic graphs bearing particular importance. We study a model which is meant to simulate the community structure frequently observed in empirical networks [141], which we term the *dynamic stochastic blockmodel*. This model is a dynamic variant of a typical community model which has been the focus of several recent works [144, 10, 32, 2].

Many previous approaches to understanding the evolution of community structure focus on first decomposing the graph into communities. In order to circumvent the cost of such a decomposition, we utilize a distance function that performs pairwise comparison of graphs at timesteps t and $t + \Delta t$ without needing any direct mapping of the community structure. In particular, we use the *resistance perturbation distance*, originally proposed in [107]. The RP-distance is a true metric, which allows for rigorous analysis not typically possible with similarity measures (for an example of a similarity measure for pairwise graph comparison, see [84]).

We perform a detailed analysis of the ability of the RP-distance to detect anomalous evolution in the dynamic stochastic blockmodel. The model is formed by sequentially adding vertices to the graph, and randomly connecting them within and across their assigned community In particular, we determine in which asymptotic regimes of in-community and cross-community density the RP-

distance can detect the creation of cross-community edges. The stochastic nature of this model suggests that our challenge is to detect the addition of a cross-community edge within the normal fluctuations present in the model. This work led to [157].

This analysis of the resistance perturbation distance motivates a more general study of the performance of various graph distance functions in detecting various structural differences. Such distance functions are applicable not only in detecting anomalous evolution of a dynamic graph, but also in determining whether a graph belongs to a given population, for example in problems of medical diagnosis in neuroscience [139]. Although many graph distance tools perform comparisons to other methods as part of their development [107, 84], these comparisons are often limited in their scope, and not sufficiently general to inform decisions on a variety of application areas. Surveys such as [6] address the academic understanding of a variety of methods present in the literature, but also fail to provide the empirical testing necessary to properly understand the applicability of various methods.

In Chapter 5, we perform a large study of graph distances, in application to both synthetic and empirical data. We expand our scope to allow both true metrics and pseudo-distances, which may not be injective. The study focuses on metrics which can be calculated efficiently, as we anticipate applications on graphs of the size of 10^5 and above. The use of random graph models allows us to understand how each distance responds to each structural signature and scale present in the models. We are particularly interested in the typical structural scale at which each distance is most responsive; for example, the graph edit distance is naturally myopic in scale, whereas the first few eigenvalues of the adjacency matrix naturally describe global community structure [91].

We apply the insights gleaned in this way to the analysis of empirical graph data. In particular, we analyze a social contact graph [138] to detect the topological changes encountered in transitions between activities (lunch, class, recess) and also examine neurological connectivity graphs in order to attempt to classify subject with and without autism spectrum disorder [1].

The structure of this thesis is an acknowledgement of the compartmentalized nature of the subject matter. In each chapter we provide an introduction, conclusion, and appendices as appropriate. These chapters are largely self-contained, and can be read without reference to one another, although the material in Chapter 5 is strongly connected to that in Chapter 4. Finally, we provide a conclusion which summarizes the important results of each work, and suggest directions in which each research program could evolve.

Chapter 2

Performance of Test Supermartingale Confidence Intervals for the Success Probability of Bernoulli Trials

This chapter written in collaboration with Emanuel Knill, [1] [2] Kevin Coakley, [3] and Yanbao Zhang. [4] [5]

2.1 Introduction

Experiments in physics require very high confidence to justify claims of discovery or to unambiguously exclude alternative explanations [43]. Particularly striking examples in the foundations of physics are experiments to demonstrate that theories based on local hidden variables, called local realist (LR) theories, cannot explain the statistics observed in quantum experiments called Bell tests. See Ref. [54] for a review and Refs. [65, 131, 57, 125] for the most definitive experiments to date. Successful Bell tests imply the presence of some randomness in the observed statistics. As a result, one of the most notable applications of Bell tests is to randomness generation [4]. In this application, it is necessary to certify the randomness generated, and these certificates are equivalent to extremely small significance levels in an appropriately formulated hypothesis test. In general, such extreme significance levels are frequently required in protocols for communication or computation to ensure performance.

[1] National Institute of Standards and Technology, Boulder, Colorado 80305, USA
[2] Center for Theory of Quantum Matter, University of Colorado, Boulder, Colorado 80309, USA
[3] National Institute of Standards and Technology, Boulder, Colorado 80305, USA
[4] Institute for Quantum Computing and Department of Physics and Astronomy, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[5] NTT Basic Research Laboratories, NTT Corporation, 3-1 Morinosato-Wakamiya, Atsugi, Kanagawa 243-0198, Japan

Bell tests consist of a sequence of "trials", each of which gives a result M_i. LR models restrict the statistics of the M_i and therefore constitute a composite null hypothesis to be rejected. Traditionally, data has been analyzed by estimating the value of a Bell function $\hat{B}$ and its standard error $\hat{\sigma}$ from the collective result statistics (see [162, 89]). Under the null hypothesis, $\hat{B}$ is expected to be negative, so a large value of $\hat{B}$ compared to $\hat{\sigma}$ is considered strong evidence against the null hypothesis. This method suffers from several problems, including the failure of the Gaussian approximation in the extreme tails and the fact that the trials are observably not independent and identically distributed (i.i.d.) [162].

In Ref. [162] a method was introduced that can give rigorous p-value bounds against LR. These p-value bounds are memory-robust, that is, without any assumptions on dependence of trial statistics on earlier trials. The method can be seen as an application of test supermartingales as defined in Ref. [130]. Test supermartingales were first considered, and many of their basic properties were proved, by Ville [148] in the same work that introduced the notion of martingales. The method involves constructing a non-negative stochastic process V_i determined by $(M_j)_{j\leq i}$ such that the initial value is $V_0 = 1$ and, under LR models, the expectations conditional on all past events are non-increasing. As explained further below, the final value of V_i in a sequence of n trials has expectation bounded by 1, so its inverse $p = 1/V_n$ is a p-value bound according to Markov's inequality. A large observed value of such a test supermartingale thus provides evidence against LR models. Refs. [162, 163] give methods to construct V_i that achieve asymptotically optimal gain rate $\mathbb{E}(-\log(p)/n)$ for i.i.d. trials, where $\mathbb{E}(\ldots)$ is the expectation functional. This is typically an improvement over other valid memory-robust Bell tests. Additional benefits are that V_i can be constructed adaptively based on the observed statistics, and the p-value bounds remain valid even if the experiment is stopped based on the current value of V_i. These techniques were successfully applied to experimental data from a Bell test with photons where other methods fail [33].

Although the terminology is apparently relatively recent, test supermartingales have traditionally played a major theoretical role. Carefully constructed test supermartingales contribute to the asymptotic analysis of distributions and proofs of large deviation bounds. They can be con-

structed for any convex-closed null hypothesis viewed as a set of distributions, so they can be used for memory- and stopping-robust adaptive hypothesis tests in some generality. The application to Bell tests shows that at least in a regime where high significance results are required, test supermartingales can perform well or better than other methods. Here we compare the performance of test supermartingales directly to (1) the standard large deviation bounds based on the Chernoff-Hoeffding inequality [31, 69], and (2) "exact" p-value calculations. Our comparison is for a case where all calculations can be performed efficiently, namely for testing the success probability in Bernoulli trials. The three p-value bounds thus obtained have asymptotically optimal gain rates. Not surprisingly, for any given experiment, test supermartingales yield systematically worse p-value bounds, but the difference is much smaller than the experiment-to-experiment variation. This effect can be viewed as the cost of robustness against arbitrary stopping rules. For ease of calculation, we do not use an optimal test supermartingale construction, but we expect similar results no matter which test supermartingale is used.

Any hypothesis test parametrized by ϕ can be used to construct confidence regions for ϕ by acceptance region inversion (see Ref. [133], Sect. 7.1.2). Motivated by this observation, we consider the use of test supermartingales for determining confidence regions. We expect that they perform well in the high-confidence regime, with an increase in region size associated with robustness against stopping rules. We therefore compared the methods mentioned above for determining confidence intervals for the success probability in Bernoulli trials. After normalizing the difference between the interval endpoints and the success probability by the standard deviation, which is $O(1/\sqrt{n})$, we find that while large deviation bounds and exact regions differ by a constant at fixed confidence levels, the test supermartingale's normalized endpoint deviation is $\Omega(\sqrt{\log(n)})$ instead of $O(1)$. This effect was noted in Ref. [130] and partially reflects a suboptimal choice of supermartingale. To maintain robustness against stopping rule, one expects $\Omega(\sqrt{\log\log(n)})$ according to the law of the iterated logarithm. However, we note that if the number of trials n is fixed in advance, the normalized endpoint deviation can be reduced to $O(1)$ with an adaptive test supermartingale. So although the increase in confidence region necessitated by stopping rule robustness is not so large for reasonably

sized n, when n is known ahead of time it can, in principle, be avoided without losing the ability to adapt the test supermartingale on the fly during the experiment in non-i.i.d. situations.

The remainder of the chapter is structured as follows. We establish the notation to be used and define the basic concepts in Sect. 2.2. Here we also explain how adaptivity can help reject hypotheses for stochastic processes. We introduce the three methods to be applied to Bernoulli trials in Sect. 2.3. Here we also establish the basic monotonicity properties and relationships of the three p-value bounds obtained. In Sect. 2.4 we determine the behavior of the p-value bounds in detail, including their asymptotic behavior. In Sect. 2.5 we introduce the confidence intervals obtained by acceptance region inversion. We focus on one-sided intervals determined by lower bounds but note that the results apply to two-sided intervals. The observations in Sects. 2.4 and 2.5 are based on theorems whose proofs can be found in the Appendix. While many of the observations in these sections can ignore asymptotically small terms, the results in the Appendix uncompromisingly determine interval bounds for all relevant expressions, with explicit constants. Concluding remarks can be found in Sect. 2.6.

2.2 Basic Concepts

We use the usual conventions for random variables (RVs) and their values. RVs are denoted by capital letters such as $X, Y, \ldots$ and their values by the corresponding lower case letters $x, y, \ldots$. All our RVs are finite valued. Probabilities and expectations are denoted by $\mathbb{P}(\ldots)$ and $\mathbb{E}(\ldots)$, respectively. For a formula ϕ, the expression $\{\phi\}$ refers to the event where the formula is true. The notation $\mu(X)$ refers to the distribution of X induced on its space of values. We use the usual conventions for conditional probabilities and expectations. Also, $\mu(X|\phi)$ denotes the probability distribution induced by X conditional on the event $\{\phi\}$.

We consider stochastic sequences of RVs such as $\mathbf{X} = (X_i)_{i=1}^n$ and $\mathbf{X}_{\leq k} = (X_i)_{i=1}^k$. We think of the X_i as the outcomes from a sequence of trials. For our study, we consider $\mathbf{B} = (B_i)_{i=1}^n$, where the B_i are $\{0,1\}$-valued RVs. The standard $\{0,1\}$-valued RV with parameter θ is the Bernoulli RV B satisfying $\mathbb{E}(B) = \theta$. The parameter θ is also referred to as the success probability. We

denote the distribution of B by ν_θ. We define $S_k = \sum_{i=1}^k B_i$ and $\hat{\Theta}_k = S_k/k$. We extend the RV conventions to the Greek letter $\hat{\Theta}_k$. That is, $\hat{\theta}_k = s_k/k = \sum_{i=1}^k b_i/k$ is the value of the RV $\hat{\Theta}_k$ determined by the values b_i of B_i. We may omit subscripts on statistics such as S_n and $\hat{\Theta}_n$ when they are based on the full set of n samples. Some expressions involving $\hat{\Theta}_n$ require that $n\hat{\Theta}_n$ is an integer, which is assured by the definition.

A null hypothesis for X is equivalent to a set $\mathcal{H}_0$ of distributions of X, which we refer to as the "null". For our study of Bernoulli RVs, we consider the nulls

$$\mathcal{B}_\varphi = \{\nu_\theta | \theta \leq \varphi\} \tag{2.1}$$

parametrized by $0 \leq \varphi \leq 1$. the set of distributions of Bernoulli RVs with $\mathbb{P}(B = 1) \leq \varphi$. One can test the null hypothesis determined by a null by means of special statistics called p-value bounds. A statistic $P = P(X) \geq 0$ is a p-value bound for $\mathcal{H}_0$ if for all $\mu \in \mathcal{H}_0$ and $p \geq 0$, $\mathbb{P}_\mu(P \leq p) \leq p$. Here, the subscript μ on $\mathbb{P}_\mu(\ldots)$ indicates the distribution with respect to which the probabilities are to be calculated. We usually just write "p-value" instead of "p-value bound", even when the bounds are not achieved by a member of $\mathcal{H}_0$. Small p-values are strong evidence against the null. Since we are interested in very small p-values, we preferentially use their negative logarithm $-\log(P)$ and call this the $\log(p)$-value. In this work, logarithms are base e by default.

A general method for constructing p-values is to start with an arbitrary real-valued RV Q jointly distributed with X. Usually Q is a function of X. Define the worst-case tail probability of Q as $P(q) = \sup_{\mu \in \mathcal{H}_0} \mathbb{P}_\mu(Q \geq q)$. Then $P(Q)$ is a p-value for $\mathcal{H}_0$. The argument is standard. Define $F_\mu(q) = \mathbb{P}_\mu(Q \geq q)$. The function F_μ is non-increasing. We need to show that for all $\mu \in \mathcal{H}_0$, $\mathbb{P}_\mu(P(Q) \leq p) \leq p$. Since $F_\mu(q) \leq P(q)$, we have $\mathbb{P}_\mu(P(Q) \leq p) \leq \mathbb{P}_\mu(F_\mu(Q) \leq p)$. The set $\{q : F_\mu(q) \leq p\}$ is either of the form $[q_{\min}, \infty)$ or $(q_{\min}, \infty)$ for some $q_{\min}$. In the first case, $\mathbb{P}_\mu(F_\mu(Q) \leq p) = \mathbb{P}_\mu(Q \geq q_{\min}) = F_\mu(q_{\min}) \leq p$. In the second, $\mathbb{P}_\mu(F_\mu(Q) \leq p) = \mathbb{P}_\mu\left(\bigcup_n \{q : q \geq q_{\min} + 1/n\}\right) = \lim_n \mathbb{P}_\mu(\{q : q \geq q_{\min} + 1/n\}) = \lim_n \mathbb{P}_\mu(F_\mu(Q) \leq q_{\min} + 1/n) \leq p$, with σ-additivity applied to the countable monotone union.

When referring to $\mathcal{H}_0$ as a null for $\mathbf{X}$, we mean that $\mathcal{H}_0$ consists of the distributions where

the X_i are i.i.d., with X_i distributed according to μ for some fixed μ independent of i. To go beyond i.i.d., we extend $\mathcal{H}_0$ to the set of distributions of $\mathbf{X}$ that have the property that for all $\mathbf{x}_{\leq i-1}$, $\mu(X_i|\mathbf{X}_{\leq i-1} = \mathbf{x}_{\leq i-1}) = \mu_i$ for some $\mu_i \in \mathcal{H}_0$, where μ_i depends on i and $\mathbf{x}_{\leq i-1}$. We denote the extended null by $\overline{\mathcal{H}_0}$. In particular,

$$\overline{\mathcal{B}_\varphi} = \{\mu : \text{for all } i \text{ and } \mathbf{b}_{\leq i-1},\ \mu(B_i|\mathbf{B}_{\leq i-1} = \mathbf{b}_{\leq i-1}) = \nu_\theta \text{ for some } \theta \leq \varphi\}. \tag{2.2}$$

The LR models mentioned in the introduction constitute a particular null $\mathcal{H}_{\text{LR}}$ for sequences of trials called Bell tests. In Ref. [162], a technique called the probability-based ratio (PBR) method was introduced to construct p-values P_n that achieve asymptotically optimal gain rates defined as $\mathbb{E}(\log(1/P_n))/n$. The method is best understood as a way of constructing a test supermartingale for $\mathcal{H}_{\text{LR}}$. A test supermartingale of $\mathbf{X}$ for $\mathcal{H}_0$ is a stochastic sequence $\mathbf{T} = (T_i)_{i=0}^n$ where T_i is a function of $\mathbf{X}_{\leq i}$, $T_0 = 1$, $T_i \geq 0$ and for all $\mu \in \mathcal{H}_0$, $\mathbb{E}_\mu(T_{i+1}|\mathbf{X}_{\leq i}) \leq T_i$. In this work, to avoid unwanted boundary cases, we further require T_i to be positive. The definition of test supermartingale used here is not the most general one because we consider only discrete time and avoid the customary increasing sequence of σ-algebra by making it dependent on an explicit stochastic sequence $\mathbf{X}$. Every test supermartingale defines a p-value by $P_n = 1/T_n$. This follows from $\mathbb{E}(T_n) \leq T_0 = 1$ (one of the characteristic properties of supermartingales) and Markov's inequality for non-negative statistics, according to which $\mathbb{P}(T_n \geq \kappa) \leq \mathbb{E}(T_n)/\kappa \leq 1/\kappa$. From martingale theory, the stopped process T_τ for any stopping rule τ with respect to $\mathbf{X}$ also defines a p-value by $P = 1/T_\tau$. Further, $P_n^* = 1/\max_{i=1}^n T_i$ also defines a p-value. See Ref. [130] for a discussion and examples.

A test supermartingale $\mathbf{T}$ can be viewed as the running product of the $F_i = T_i/T_{i-1}$, which we call the test factors of $\mathbf{T}$. The defining properties of $\mathbf{T}$ are equivalent to having $F_i > 0$ and $\mathbb{E}(F_i|\mathbf{X}_{\leq i-1}) \leq 1$ for all distributions in the null, for all i. The PBR method adaptively constructs F_i as a function of the next trial outcome X_i from the earlier trial outcomes $\mathbf{X}_{\leq i-1}$. The method is designed for testing $\overline{\mathcal{H}_0}$ for a closed convex null $\mathcal{H}_0$, where asymptotically optimally gain rates are achieved when the trials are i.i.d. with a trial distribution ν not in $\mathcal{H}_0$. If ν were known, the optimal test factor would be given by $x \mapsto \nu(x)/\mu(x)$, where $\mu \in \mathcal{H}_0$ is the distribution in $\mathcal{H}_0$ closest to ν

in Kullback-Leibler (KL) divergence $\mathrm{KL}(\nu|\mu) = \sum_x \nu(x) \log(\nu(x)/\mu(x))$ [87]. Since ν is not known, the PBR method obtains an empirical estimate $\hat{\nu}$ of ν from $\mathbf{x}_{\leq i-1}$ and other information available before the i'th trial. It then determines the KL-closest $\mu \in \mathcal{H}_0$ to $\hat{\nu}$. The test factor F_i is then given by $F_i(x) = \hat{\nu}(x)/\mu(x)$. The test factors satisfy $\mathbb{E}_{\mu'}(F_i) \leq 1$ for all $\mu' \in \mathcal{H}_0$, see Ref. [162] for a proof and applications to the problem of testing LR.

The ability to choose test factors adaptively helps reject extended nulls when the distributions vary as the experiment progresses, both when the distributions are still independent (so only the parameters vary) and when the parameters depend on past outcomes. Suppose that the distributions are sufficiently stable so that the empirical frequencies over the past k trials are statistically close to the next trial's probability distribution. Then we can adaptively compute the test factor to be used for the next trial from the past k trials' empirical frequencies, for example by following the strategy outlined in the previous paragraph. The procedure now has an opportunity to reject an extended null provided only that there is a sufficiently long period where the original null does not hold. For example, consider the extended null $\overline{\mathcal{B}_\varphi}$. The true success probabilities θ_i at the i'th trial may vary, maybe as a result of changes in experimental parameters that need to be calibrated. Suppose that the goal is to calibrate for $\theta_i > \varphi$. If we use adaptive test factors and find at some point that we cannot reject $\overline{\mathcal{B}_\varphi}$ according to the running product of the test factors, we can recalibrate during the experiment. If the the recalibration succeeds at pushing θ_i above φ for the remaining trials, we may still reject the extended null by the end of the experiment. In many cases, the analysis is performed after the experiment, or it may not be possible to stop the experiment for recalibration. For this situation, if the frequencies for a run of k trials clearly show that $\theta_i < \varphi$, the adaptive test factors chosen would tend to be trivial (equal to 1), in which case the next trials do not contribute to the final test factor product. This is in contrast to a hypothesis test based on the final sum of the outcomes for which all trials contribute equally.

Let φ be a parameter of distributions of X. Here, φ need not determine the distributions. There is a close relationship between methods for determining confidence sets for φ and hypothesis tests. Let $\mathcal{H}_\varphi$ be a null such that for all distributions μ with parameter φ, $\mu \in \mathcal{H}_\varphi$. Given a family

of hypothesis tests with p-values P_φ for $\mathcal{H}_\varphi$, we can construct confidence sets for φ by inverting the acceptance region of P_φ, see Ref. [133], Sect. 7.1.2. According to this construction, the confidence set C_a at level a is given by $\{\varphi | P_\varphi(X) \geq a\}$ and is a random quantity. The defining property of a level a confidence set is that its coverage probability satisfies $\mathbb{P}_\mu(\varphi \in C_a) \geq 1-a$ for all distributions $\mu \in \mathcal{H}_\varphi$. When we use this construction for sequences $\mathbf{B}$ of i.i.d. Bernoulli RVs with the null $\mathcal{B}_\varphi$, we obtain one-sided confidence intervals of the form $[\varphi_0, 1]$ for $\theta = \mathbb{E}(B_i)$. When the confidence set is a one-sided interval of this type, we refer to φ_0 as the confidence lower bound or endpoint. If $\mathbf{B}$ has a distribution μ that is not necessarily i.i.d., we can define $\Theta_{\max} = \max_{i \leq n} \mathbb{E}_\mu(B_i | \mathbf{B}_{\leq i-1})$. If we use acceptance region inversion with the extended null $\overline{\mathcal{B}_\varphi}$, we obtain a confidence region for $\Theta_{\max}$. Note that $\Theta_{\max}$ is an RV whose value is covered by the confidence set with probability at least $1-a$. The confidence set need not be an interval in general, but including everything between its infimum and its supremum increases the coverage probability, so the set can be converted into an interval if desired.

While our focus is on one-sided confidence intervals, our observations immediately apply to two-sided intervals ones with a standard method of obtaining a two-sided confidence interval from two one-sided intervals. For our example, we can obtain confidence upper bounds at level a by symmetry, for example by relabeling the Bernoulli outcomes $0 \mapsto 1$ and $1 \mapsto 0$. To obtain a two-sided interval at level a, we compute lower and upper bounds at level $a/2$. The two-sided interval is the interval between the bounds. The coverage probability of the two-sided interval is valid according to the union bound applied to maximum non-coverage probabilities of the two one-sided intervals.

2.3 Bernoulli Hypothesis Tests

We compare three hypothesis tests for the nulls $\mathcal{B}_\varphi$ or the extended nulls $\overline{\mathcal{B}_\varphi}$: The "exact" test with p-value P_{X}, the Chernoff-Hoeffding test with p-value P_{CH} and a PBR test with p-value P_{PBR}. In discussing properties of these tests with respect to the hypothesis parameter φ, the true success probability θ and the empirical success probability $\hat{\Theta}$, we generally assume that these parameters

are in the interior of their range. In particular, $0 < \varphi < 1$, $0 < \theta < 1$, and $0 < \hat{\Theta} < 1$. When discussing purely functional properties with respect to values $\hat{\theta}$ of $\hat{\Theta}$, we use the variable t instead of $\hat{\theta}$. By default nt is a positive integer.

The p-value for the exact test is obtained from the tail for i.i.d. Bernoulli RVs:

$$P_{\mathrm{X},n}(\hat{\Theta}|\varphi) = \sum_{k \geq \hat{\Theta} n} \binom{n}{k} \varphi^k (1-\varphi)^{n-k}, \tag{2.3}$$

where $\hat{\Theta} = S_n/n = \sum_{i=1}^n B_i/n$ as defined in Sect. 2.2. Note that unlike the other p-values we consider, $P_{\mathrm{X},n}$ is not just a p-value bound. It is achieved by a member of the null. The quantity $P_{\mathrm{X},n}(t|\varphi)$ is decreasing as a function of t, given $0 < \varphi < 1$. It is smooth and monotonically increasing as a function of φ, given $t > 0$. To see this, compute

$$\begin{aligned}
\frac{d}{d\varphi} P_{\mathrm{X},n}(t|\varphi) &= \sum_{i=nt}^{n} \varphi^i (1-\varphi)^{n-i} \binom{n}{i} \left(\frac{i}{\varphi} - \frac{n-i}{1-\varphi} \right) \\
&= n \sum_{i=nt}^{n} \varphi^{i-1} (1-\varphi)^{n-i} \binom{n-1}{i-1} - n \sum_{i=nt}^{n-1} \varphi^i (1-\varphi)^{n-i-1} \binom{n-1}{i} \\
&= n \left(\sum_{i=nt-1}^{n-1} \varphi^i (1-\varphi)^{n-1-i} \binom{n-1}{i} - \sum_{i=nt}^{n-1} \varphi^i (1-\varphi)^{n-1-i} \binom{n-1}{i} \right) \\
&= n\, \varphi^{nt-1} (1-\varphi)^{n(1-t)} \binom{n-1}{nt-1}.
\end{aligned} \tag{2.4}$$

This is positive for $\varphi \in (0,1)$. The probability that $S_n \geq tn$, given that all B_i are distributed as ν_θ with $\theta \leq \varphi$, is bounded by $P_{\mathrm{X},n}(t|\theta) \leq P_{\mathrm{X},n}(t|\varphi)$. That P_{X} is a p-value for the case where the null is restricted to i.i.d. distributions now follows from the standard construction of p-values from worst-case (over the null) tails of statistics (here S_n) as explained in the previous section. That P_{X} is a p-value for the extended null $\overline{\mathcal{B}_\varphi}$ follows from the observations that the tail probabilities of S_n are linear functions of the distribution parameters $\theta_1, \theta_2, ..., \theta_n$ where $\theta_i \leq \varphi, i = 1, 2, ..., n$, the extremal distributions in $\overline{\mathcal{B}_\varphi}$ have B_i independent with $\mathbb{P}(B_i = 1) = \theta_i \leq \varphi$, and the tail probabilities of S_n are monotonically increasing in $\mathbb{P}(B_i = 1)$ for each i separately. See also Ref. [22], App. C.

Define $\hat{\Theta}_{\max} = \max(\hat{\Theta}, \varphi)$. The p-value for the Chernoff-Hoeffding test is the optimal

Chernoff-Hoeffding bound [31, 69] for a binary random variable given by

$$P_{\mathrm{CH},n}(\hat{\Theta}|\varphi) = \left(\frac{\varphi}{\Theta_{\max}}\right)^{n\Theta_{\max}} \left(\frac{1-\varphi}{1-\Theta_{\max}}\right)^{n(1-\Theta_{\max})} = \begin{cases} \left(\frac{\varphi}{\hat{\Theta}}\right)^{n\hat{\Theta}} \left(\frac{1-\varphi}{1-\hat{\Theta}}\right)^{n(1-\hat{\Theta})} & \text{if } \hat{\Theta} \geq \varphi, \\ 1 & \text{otherwise.} \end{cases} \tag{2.5}$$

This is a p-value for our setting because $P_{\mathrm{CH},n}(t|\varphi) \geq P_{\mathrm{X},n}(t|\varphi)$, see Ref. [69]. For $\varphi \leq t$, we have $-\log(P_{\mathrm{CH},n}(t|\varphi)) = n\mathrm{KL}(\nu_t|\nu_\varphi)$. We abbreviate $\mathrm{KL}(\nu_t|\nu_\varphi)$ by $\mathrm{KL}(t|\varphi)$. For $\varphi \leq t < 1$, $P_{\mathrm{CH},n}(t|\varphi)$ is monotonically increasing in φ, and decreasing in t. For $0 \leq t \leq \varphi$, it is constant.

The p-value for the PBR test that we use for comparison is constructed from a p-value for the point null $\{\nu_\varphi\}$ defined as

$$P^0_{\mathrm{PBR},n}(\hat{\Theta}|\varphi) = \varphi^{n\hat{\Theta}}(1-\varphi)^{n(1-\hat{\Theta})}(n+1)\binom{n}{n\hat{\Theta}}. \tag{2.6}$$

The PBR test's p-value for $\overline{\mathcal{B}_\varphi}$ is

$$P_{\mathrm{PBR},n}(\hat{\Theta}|\varphi) = \max_{0\leq\varphi'\leq\varphi} P^0_{\mathrm{PBR},n}(\hat{\Theta}|\varphi'). \tag{2.7}$$

That P_{PBR} is a p-value for $\overline{\mathcal{B}_\varphi}$ is shown below. As a function of φ, $P^0_{\mathrm{PBR},n}(t|\varphi)$ has an isolated maximum at $\varphi = t$. This can be seen by differentiating $\log\left(\varphi^t(1-\varphi)^{1-t}\right) = t\log(\varphi) + (1-t)\log(1-\varphi)$. Thus in Eq. 2.7 when $\varphi \geq \hat{\Theta}$, the maximum is achieved by $\varphi' = \hat{\Theta}$. We can therefore write Eq. 2.7 as

$$P_{\mathrm{PBR},n}(\hat{\Theta}|\varphi) = \begin{cases} P^0_{\mathrm{PBR},n}(\hat{\Theta}|\varphi) & \text{if } \hat{\Theta} \geq \varphi, \\ P^0_{\mathrm{PBR},n}(\hat{\Theta}|\hat{\Theta}) & \text{otherwise.} \end{cases} \tag{2.8}$$

By definition, $P_{\mathrm{PBR},n}(t|\varphi)$ is non-decreasing in φ and strictly increasing for $\varphi \leq t$. As a function of t, it is strictly decreasing for $t \geq \varphi$ (integer-valued nt). To see this, consider $k = nt \geq n\varphi$ and

compute the ratio of successive values as follows:

$$\begin{aligned} P^0_{\mathrm{PBR},n}((k+1)/n|\varphi)/P^0_{\mathrm{PBR},n}(k/n|\varphi) &= \frac{\varphi}{1-\varphi}\frac{n-k}{k+1} \\ &= \frac{\varphi}{1-\varphi}\frac{1-t}{t+1/n} \\ &\leq \frac{\varphi}{1-\varphi}\frac{1-t}{t} \\ &\leq 1. \end{aligned} \tag{2.9}$$

The expression for $P^0_{\mathrm{PBR},n}$ is the final value of a test supermartingale obtained by constructing test factors F_{k+1} from S_k. Define

$$\widetilde{\Theta}_k = \frac{1}{k+2}\left(S_k + 1\right). \tag{2.10}$$

Thus, $\widetilde{\Theta}_k$ would be an empirical estimate of θ if there were two initial trials B_{-1} and B_0 with values 0 and 1, respectively. The test factors are given by

$$F_{k+1}(B_{k+1}) = \left(\frac{\widetilde{\Theta}_k}{\varphi}\right)^{B_{k+1}} \left(\frac{1-\widetilde{\Theta}_k}{1-\varphi}\right)^{1-B_{k+1}}. \tag{2.11}$$

One can verify that $\mathbb{E}_{\nu_\theta}(F_{k+1}) = 1$ for $\theta = \varphi$. More generally, set $\delta = \theta - \varphi$ and compute

$$\begin{aligned} \mathbb{E}_{\nu_\theta}(F_{k+1}|\widetilde{\Theta}_k = t) &= \theta\frac{t}{\varphi} + (1-\theta)\frac{1-t}{1-\varphi} \\ &= 1 + \delta\left(\frac{t}{\varphi} - \frac{1-t}{1-\varphi}\right) \\ &= 1 + \delta\frac{t-\varphi}{\varphi(1-\varphi)}. \end{aligned} \tag{2.12}$$

As designed, $T_n = \prod_{k=1}^n F_k$ is a test supermartingale for the point null $\{\nu_\varphi\}$. Thm. 5 in App. 2.7.2, establishes that $T_n = 1/P^0_{\mathrm{PBR},n}(\hat{\Theta}|\varphi)$. The definition of $P_{\mathrm{PBR},n}(\hat{\Theta}|\varphi)$ as a maximum of p-values for $\nu_{\varphi'}$ with $\varphi' \leq \varphi$ in Eq. 2.7 ensures that $P_{\mathrm{PBR},n}(\hat{\Theta}|\varphi)$ is a p-value for $\mathcal{B}_\varphi$.

To show that P_{PBR} is a p-value for $\overline{\mathcal{B}_\varphi}$, we establish that for all t (integer-valued nt), $P_{\mathrm{PBR},n}(t|\varphi) \geq P_{\mathrm{CH},n}(t|\varphi)$. By direct calculation for both $\varphi \leq t$ and $t \leq \varphi$, we have

$$P_{\mathrm{PBR},n}(t|\varphi)/P_{\mathrm{CH},n}(t|\varphi) = t^{nt}(1-t)^{n(1-t)}(n+1)\binom{n}{nt}. \tag{2.13}$$

The expression $t^k(1-t)^k\binom{n}{k}$ is maximized at $k = nt$ as can be seen by considering ratios for successive values of k and the calculation in Eq. 2.9, now applied also for $k < nt$. Therefore,

$$t^{nt}(1-t)^{n(1-t)}(n+1)\binom{n}{nt} = \sum_{k=0}^{n} t^{nt}(1-t)^{n(1-t)}\binom{n}{nt} \geq \sum_{k=0}^{n} t^k(1-t)^k\binom{n}{k} = 1. \tag{2.14}$$

A better choice for test factors to construct a test supermartingale to test $\overline{\mathcal{B}_\varphi}$ would be

$$T'_{k+1} = \begin{cases} T_{k+1} & \text{if } \widetilde{\Theta}_k \geq \varphi, \\ 1 & \text{otherwise.} \end{cases} \tag{2.15}$$

This choice ensures that $\mathbb{E}_{\nu_\theta}(F_{k+1}|\mathbf{B}_{\leq k}) \leq 1$ for all $\theta \leq \varphi$ but the final value of the test supermartingale obtained by multiplying these test factors is not determined by S_n, which would complicate our study.

We summarize the observations about the three tests in the following theorem.

Theorem 1. *We have*

$$P_{\mathrm{X}} \leq P_{\mathrm{CH}} \leq P_{\mathrm{PBR}}. \tag{2.16}$$

The three tests satisfy the following monotonicity properties for $0 < \varphi < 1$ *and* $0 < t < 1$ *with* nt *integer-valued:*

$P_{\mathrm{X}}(t|\varphi)$ is strictly increasing in φ and strictly decreasing as a function of t.

$P_{\mathrm{CH}}(t|\varphi)$ is strictly increasing in φ for $\varphi \leq t$, constant in φ for $\varphi \geq t$, strictly decreasing in t for $t \geq \varphi$ and constant in t for $t \leq \varphi$.

$P_{\mathrm{PBR}}(t|\varphi)$ is strictly increasing in φ for $\varphi \leq t$, constant in φ for $\varphi \geq t$ and strictly decreasing in t for $t \geq \varphi$.

2.4 Comparison of p-Values

We begin by determining the relationships between P_{X}, P_{CH} and P_{PBR} more precisely. Since we are interested in small p-values, it is convenient to focus on the $\log(p)$-values instead and determine their differences to $O(1/\sqrt{n})$. Because of the identity $-\log(P_{\mathrm{CH},n}(t,\varphi)) = n\mathrm{KL}(t|\varphi)$, we

reference all $\log(p)$-values to $-\log(P_{\mathrm{CH},n})$. Here we examine the differences for $t \geq \varphi$ determined by the following theorem:

Theorem 2. *For* $0 < \varphi < t < 1$,

$$-\log(P_{\mathrm{PBR},n}(t|\varphi)) = -\log(P_{\mathrm{CH},n}(t|\varphi)) - \frac{1}{2}\log(n+1) + \frac{1}{2}\log(2\pi t(1-t)) + O\left(\frac{1}{n}\right), \tag{2.17}$$

$$-\log(P_{\mathrm{X},n}(t|\varphi)) = -\log(P_{\mathrm{CH},n}(t|\varphi)) + \frac{1}{2}\log(n) - \log\left(\sqrt{\frac{t}{2\pi(1-t)}}\frac{1-\varphi}{t-\varphi}\right) + O\left(\frac{1}{n}\right). \tag{2.18}$$

The theorem follows from Thms. 6, 7 and Cor. 8 proven in the Appendix, where explicit interval expressions are obtained for these $\log(p)$-value differences. The order notation assumes fixed $t > \varphi$. The bounds are not uniform, see the expressions in the appendix for details.

The most notable observation is that there are systematic gaps of $\log(n)/2+O(1)$ between the $\log(p)$-values. As we already knew, there is no question that the exact test is the best of the three for this simple application. While these gaps may seem large on an absolute scale, representing factors close to $\sqrt{n}$, they are in fact much smaller than the experiment-to-experiment variation of the p-values. To determine this variation, we consider the asymptotic distributions. We can readily determine that the $\log(p)$-values are asymptotically normal with standard deviations proportional to $\sqrt{n}$, which is transferred from the variance of $\hat{\Theta}$. Compared to these standard deviations the gaps are negligible. The next theorem determines the specific way in which asymptotic normality holds. Let $N(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2. The notation $X_n \xrightarrow{D} N(\mu, \sigma^2)$ means that X_n converges in distribution to the normal distribution with mean μ and variance σ^2.

Theorem 3. *Assume* $0 < \varphi < \theta < 1$. *For* $P = P_{\mathrm{CH},n}$, $P = P_{\mathrm{PBR},n}$ *or* $P = P_{\mathrm{X},n}$, *the* $\log(p)$*-value* $-\log(P)$ *converges in distribution according to*

$$\sqrt{n}(-\log(P)/n - \mathrm{KL}(\theta|\varphi)) \xrightarrow{D} N(0, \sigma_G^2), \tag{2.19}$$

with

$$\sigma_G^2 = \theta(1-\theta)\left(\log\left(\frac{\theta}{1-\theta}\frac{1-\varphi}{\varphi}\right)\right)^2.$$

The theorem is proven in the Appendix, see Thm. 10. For the rest of the chapter, we write P or P_n for the p-values of any one of the tests when it does not matter which one.

We display the behavior described in the above theorems for $n = 100$ and $\theta = 0.5$ in Fig. 2.1. We conclude that the phenomena discussed above are already apparent for small numbers of trials. For Fig. 2.1, we computed the quantiles of the $\log(p)$-values numerically using the formulas provided in the previous section, substituting for t the corresponding quantile of $\hat{\Theta}$ given that $\mathbb{P}(B = 1) = \theta$. To be explicit, let $t_{r,n}(\theta)$ be the r-quantile of $\hat{\Theta}$ defined as the minimum value $\hat{\theta}$ of $\hat{\Theta}$ satisfying $\mathbb{P}(\hat{\Theta} \leq \hat{\theta}) \geq r$. (For simplicity we do not place the quantile in the middle of the relevant gap in the distribution.) For example, $t_{0.5,n}(\theta)$ is the median. Then, by the monotonicity properties of the tests, the r-quantile of $-\log(P_n(\hat{\Theta}|\varphi))$ is given by $-\log(P_n(t_{r,n}(\theta)|\varphi))$.

As noted above, the gaps between the $\log(p)$-values are of the form $\log(n)/2 + O(1)$. In fact, it is possible to determine the asymptotic behavior of these gaps. After accounting for the explicitly given $O(1)$ terms in Thm. 2, they are asymptotically normal with variances of order $O(1/n)$. The standard deviations of the gaps are therefore small compared to their size. The precise statement of their asymptotic normality is Thm. 11 in the Appendix. Fig. 2.2 shows how these gaps depend on the value $\hat{\theta}$ of $\hat{\Theta}$ given φ. The gaps are scaled by $\log(n)$ so that they can be compared to $\log(n)/2$ visually for different values of n. The deviation from $\log(n)/2$ is most notable near the boundaries, where convergence is also slower, particularly for P_{X}. This behavior is consistent with the divergences as t approaches φ in the explicit interval bounds in Thm. 7 and Cor. 8.

2.5 Comparison of Confidence Intervals

Let P be one of $P_{\mathrm{CH},n}$, $P_{\mathrm{PBR},n}$ or $P_{\mathrm{X},n}$. Given a value $\hat{\theta}$ of $\hat{\Theta}$, the level-a confidence set determined by the test with p-value P is $I = \{\varphi | P(\hat{\theta}|\varphi) \geq a\}$. By the monotonicity properties of P, the closure of I is an interval $[\varphi_a(\hat{\theta}; P), 1]$. We can compute the endpoint φ_a by numerically inverting the exact expressions for P. An example is shown in Fig. 2.3, where we show the endpoints according to each test for $a = 0.01$ and $\hat{\theta} = 0.5$ as a function of n. All tests' endpoints converge to 0.5 as the number of trials grows. Notably, the relative separation between the endpoints is not

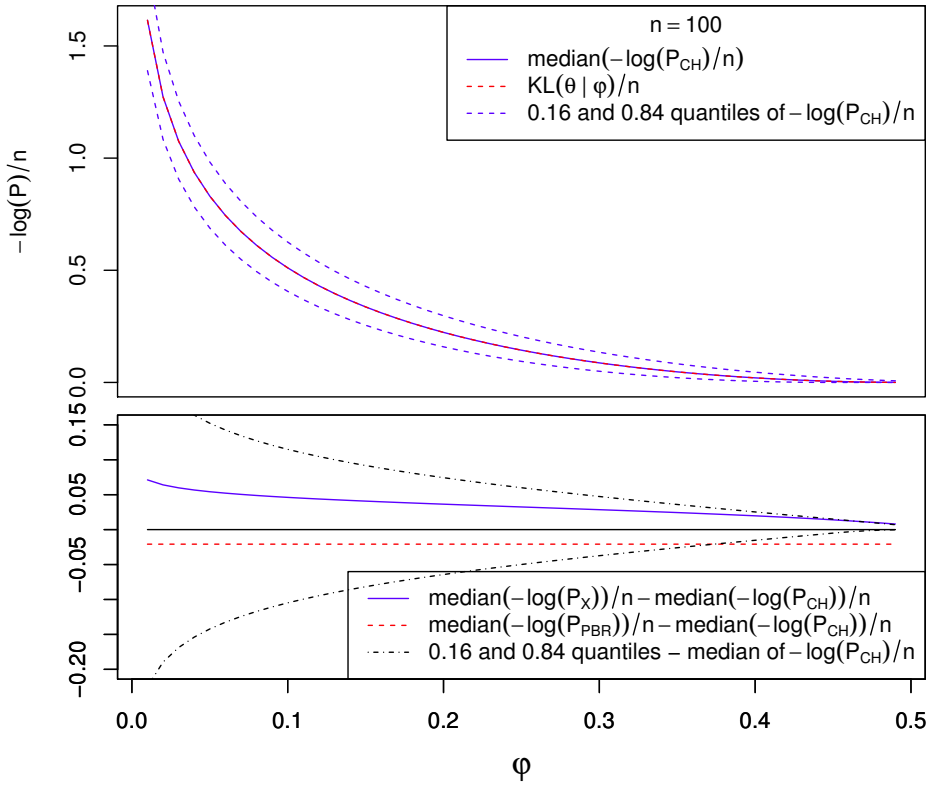

Figure 2.1: Comparison of $\log(p)$-values at $n = 100$ and $\theta = 0.5$. The top half of the figure shows the median, the 0.16 and the 0.84 quantile of $-\log(P_{\mathrm{CH},n}(\hat{\Theta}|\varphi))/n$. For $\theta = 0.5$, the median agrees with $\mathrm{KL}(\theta|\varphi)$ by symmetry. The lower half shows the median differences $-\log(P(\hat{\Theta}|\varphi))/n + \log(P_{\mathrm{CH},n}(\hat{\Theta}|\varphi))/n$ for $P = P_{\mathrm{PBR},n}$ and $P = P_{\mathrm{X},n}$. The difference between the 0.16 and 0.84 quantiles and the median for $-\log(P_{\mathrm{CH},n}(\hat{\Theta}|\varphi))/n$ are also shown where they are within the range of the plot; even for n as small as 100, they dominate the median differences, except where φ approaches $\theta = 0.5$, where the absolute p-values are no longer extremely small.

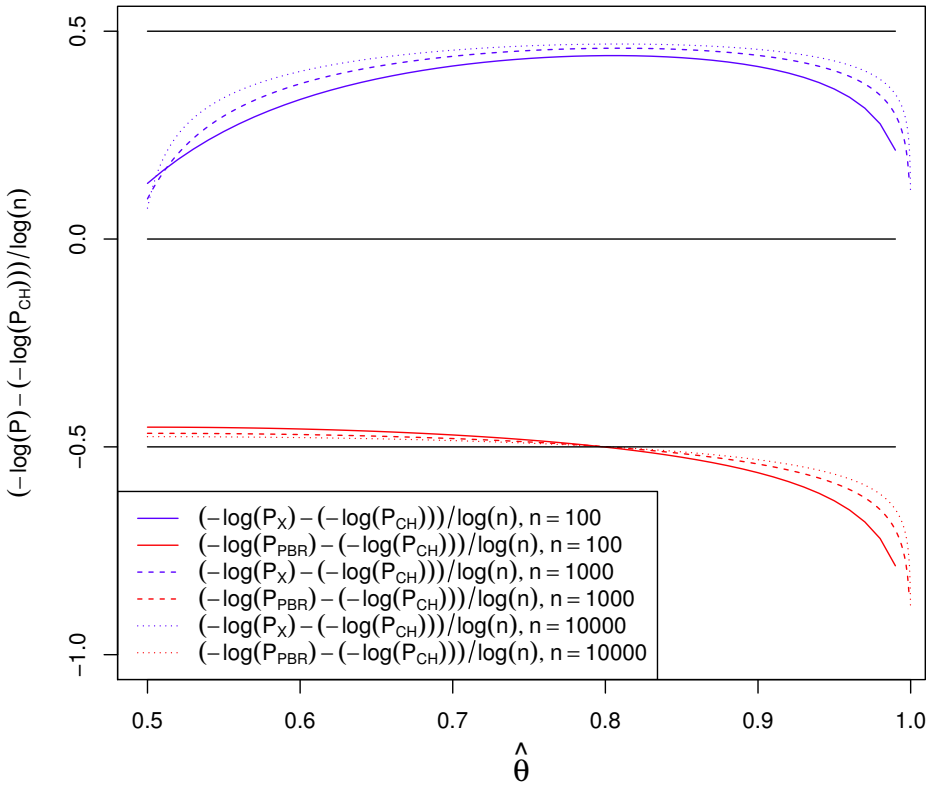

Figure 2.2: Gaps between the $\log(p)$-values depending on $\hat{\theta}$ at $\varphi = 0.5$. We show the normalized differences $\left(-\log(P_n(\hat{\theta}|\varphi)) + \log(P_{\mathrm{CH},n}(\hat{\theta}|\varphi))\right)/\log(n)$ for $P = P_{\mathrm{CH}}$ and $P = P_{\mathrm{X}}$ at $n = 100, 1000, 10000$. For large n, at constant $\hat{\theta}$ with $0.5 < \hat{\theta} < 1$, the PBR test's normalized difference converges to -0.5, and the exact test's converges to 0.5. The horizontal lines at ± 0.5 indicate this limit. The lowest order normalized asymptotic differences from ± 0.5 are $O(1/\log(n))$ and diverge at $\hat{\theta} = 0.5$ and $\hat{\theta} = 1$.

large at level $a = 0.01$.

To quantify the behavior of the endpoints for the different tests, we normalize by the empirical standard deviation $\hat{\sigma} = \sqrt{\hat{\theta}(1-\hat{\theta})/n}$. The empirical endpoint deviation is then defined as

$$\gamma_a(\hat{\theta}; P) = \frac{\hat{\theta} - \varphi_a(\hat{\theta}; P)}{\hat{\sigma}}. \tag{2.20}$$

For the exact test and for large n, we expect this quantity to be determined by the tail probabilities of a standard normal distribution. That is, if the significance a is the probability that a normal RV of variance 1 exceeds κ, we expect $\gamma_a(\hat{\theta}; P_{\mathrm{X}}) \approx \kappa$.

We take the point of view that the performance of a test is characterized by the size of the endpoint deviation. If the relative size of the deviations for two tests is close to 1 then they perform similarly for the purpose of characterizing the parameter θ. Another way of comparing the intervals obtained is to consider their coverage probabilities. For our situation, the coverage probability for test P at a can be approximated by determining a' such that $\gamma_{a'}(\theta; P_{\mathrm{X}}) = \gamma_a(\theta; P)$. From Thm. 4 below, one can infer that the coverage probability is then approximately $1-a' \geq 1-a$. The coverage probabilities can be very conservative (larger than $1-a$), particularly for small a and $P = P_{\mathrm{PBR}}$.

We determined interval bounds for the empirical endpoint deviation for all three tests. The details are in App. 2.7.5. The next theorem summarizes the results asymptotically.

Theorem 4. *Let $q(x) = -\log(\mathbb{P}_{N(0,1)}(N \geq x))$ be the negative logarithm of the tail of the standard normal. Fix $\hat{\theta} \in (0,1)$. Write $\alpha = |\log(a)|$. There is a constant c (depending on $\hat{\theta}$) such that for $\alpha \in (1, cn]$, γ_a satisfies*

$$\gamma_a(\hat{\theta}; P_{\mathrm{CH}}) = \sqrt{2\alpha} + O(\alpha/\sqrt{n}), \tag{2.21}$$

$$\gamma_a(\hat{\theta}; P_{\mathrm{PBR}}) = \sqrt{2\alpha + \log(n)/2 - \log(2\pi\hat{\theta}(1-\hat{\theta}))/2} + O(\alpha/\sqrt{n}), \tag{2.22}$$

$$\gamma_a(\hat{\theta}; P_{\mathrm{X}}) = q^{-1}(\alpha) + O(\alpha/\sqrt{n}). \tag{2.23}$$

The last expression has the following approximation relevant for sufficiently large α:

$$\gamma_a(\hat{\theta}; P_{\mathrm{X}}) = \sqrt{2\alpha - \log(2\pi) - \log(2\alpha - \log(2\pi))} + O(\log(\alpha)/\alpha^{3/2}) + O(\alpha/\sqrt{n}). \tag{2.24}$$

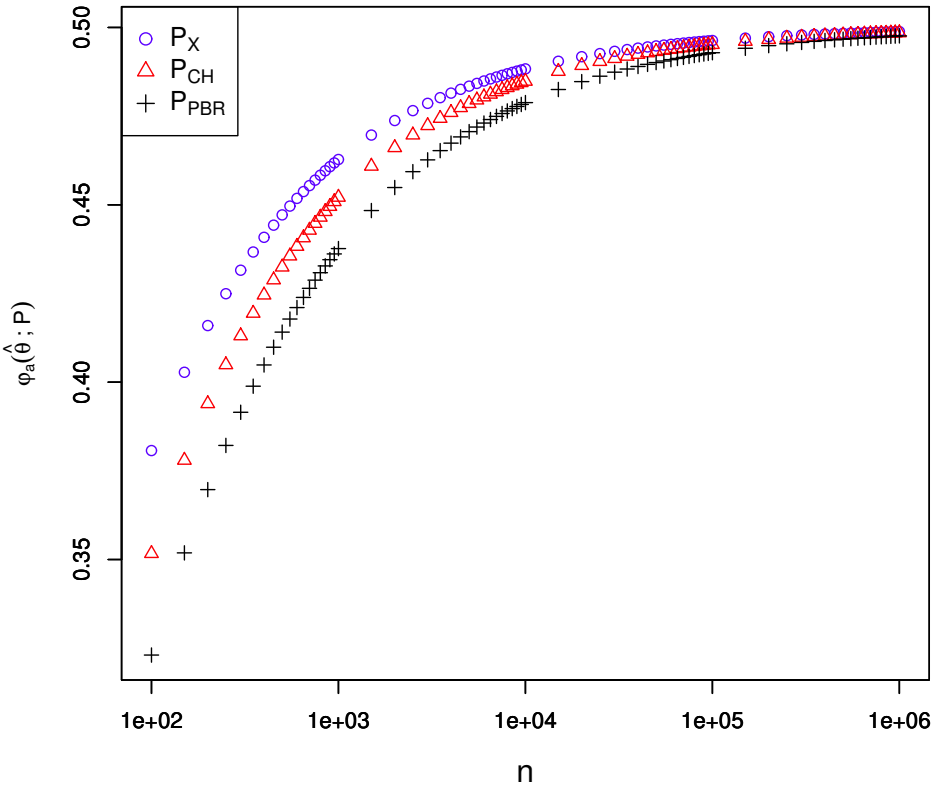

Figure 2.3: Lower endpoints for the level 0.01 confidence set of the three tests as a function of n, where $\hat{\theta} = 0.5$.

For $\alpha = o(\sqrt{n})$, the relative error of the approximation in the first two identities goes to zero as n grows. This is not the case for the last identity, where the relative error for large n is dominated by the term $O(\log(\alpha)/\alpha^{3/2})$, and large α is required for a small relative error.

Proof. The expression for $\gamma_a(\hat{\theta}; P_{\mathrm{CH}})$ is obtained from Thm. 12 in the Appendix by changing the relative approximation errors into absolute errors.

To obtain the expression for $\gamma_a(\hat{\theta}; P_{\mathrm{PBR}})$, note that the term Δ in Thm. 13 satisfies $\Delta = \log(n)/2 - \log(2\pi\hat{\theta}(1-\hat{\theta}))/2 + O(1/n)$, see Thm. 6. The $O(1/n)$ under the square root pulls out to an $O(1/(\sqrt{\max(\alpha, \log(n))n}))$ term that is dominated by $O(\alpha/\sqrt{n})$ because $\alpha \geq 1$ by assumption.

To obtain the expressions for $\gamma_a(\hat{\theta}; P_{\mathrm{X}})$, we refer to Thm. 14, where the lower bound on α implies $\alpha \geq 1 > \log(2)$. The intervals in Thm. 14 give relative errors that need to be converted to absolute quantities. By positivity and monotonicity of q^{-1}, for sufficiently large n and for some positive constants u and v, we have

$$\gamma_a(\hat{\theta}; P_{\mathrm{X}}) \in \left[q^{-1}(\alpha(1 - u\sqrt{\alpha}/\sqrt{n}))(1 - v\sqrt{\alpha}/\sqrt{n}), q^{-1}(\alpha(1 + u\sqrt{\alpha}/\sqrt{n}))(1 + v\sqrt{\alpha}/\sqrt{n})\right]. \quad (2.25)$$

Explicit values for u and v can be obtained from Thm. 14. We simplified the argument of q^{-1} by absorbing the additive terms in the theorem into the term $u\alpha\sqrt{\alpha}/\sqrt{n}$ with the constant u chosen to be sufficiently large. Consider Eq. 2.94 with $\delta = u\sqrt{\alpha}/\sqrt{n}$. For sufficiently large n, the expression in the denominator of the approximation error on the right-hand side exceeds a constant multiple of α. From this, with some new constant u',

$$\gamma_a(\hat{\theta}; P_{\mathrm{X}}) \in \left[q^{-1}(\alpha)(1 - u'\sqrt{\alpha}/\sqrt{n})(1 - v\sqrt{\alpha}/\sqrt{n}), q^{-1}(\alpha)(1 + u'\sqrt{\alpha}/\sqrt{n})(1 + v\sqrt{\alpha}/\sqrt{n})\right], \quad (2.26)$$

which, with order notation simplifies further to

$$\gamma_a(\hat{\theta}; P_{\mathrm{X}}) = q^{-1}(\alpha)(1 + O(\sqrt{\alpha}/\sqrt{n})). \quad (2.27)$$

It now suffices to apply $q^{-1}(\alpha) = O(\sqrt{\alpha})$ (see the proof of Eq. 2.24 below) and Eq. 2.23 is obtained.

For Eq. 2.24, we bound $x = q^{-1}(\alpha)$, which we can do via bounds for $\alpha = q(x)$. From the expression $q(x) = x^2/2 + \log(2\pi)/2 - \log(Y(x)) = x^2/2 + \log(2\pi)/2 + \log(x) - \log(xY(x))$ in the

statement of Thm. 14 and the bounds in Eq. 2.58, we have the two inequalities

$$q(x) = x^2/2 + \log(2\pi)/2 + \log(x) - \log(xY(x)) \geq x^2/2 + \log(2\pi)/2 + \log(x), \tag{2.28}$$

$$q(x) \leq x^2/2 + \log(2\pi)/2 + \log(x) + 1/x^2. \tag{2.29}$$

Let $l(x) = x^2/2 + \log(2\pi)/2 + \log(x)$, which is monotonically increasing, as is q. The first inequality implies that $q^{-1} \leq l^{-1}$. We need a bound of the form $q(x) \leq dx^2$, from which we can conclude that $x^2 \geq \alpha/d$. A bound of this type can be obtained from Eq. 2.91 in the Appendix. For definiteness, we restrict to $\alpha \geq 6$ and show that the bound holds with $d = 1$. By Eq. 2.29, it suffices to establish that for $x \geq \sqrt{6}$, $l(x) + 1/x^2 \leq x^2$. Since $\log(2\pi)/2 \leq 1$, we have $\log(2\pi)/2 + \log(x) + 1/x^2 \leq 1 + \log(1 + (x-1)) + 1/x^2 \leq x + 1/x^2$. For $x \geq 9/4$, $x + 1/x^2 \leq x^2/2$. To finish the argument, apply the inequality $\sqrt{6} \geq 9/4$.

Given the bound $x^2 \geq \alpha$, Eq. 2.29 becomes $q(x) = \alpha \leq l(x) + 1/\alpha$. With Eq. 2.28 we get $\alpha = q(x) \in l(x) + [0,1]/\alpha$. Equivalently,

$$l(x) \in \alpha + \frac{1}{\alpha}[-1, 0]. \tag{2.30}$$

Applying the monotone l^{-1} on both sides gives

$$x = l^{-1}(l(x)) \in l^{-1}\left(\alpha + \frac{1}{\alpha}[-1, 0]\right). \tag{2.31}$$

Let α' satisfy $x = l^{-1}(\alpha')$ with $\alpha' = \alpha + \delta$ and $\delta \in [-1, 0]/\alpha$. Write $z = x^2$ and $c = \log(2\pi)$. We have $l(z^{1/2}) = z/2 + c/2 + \log(z)/2 = \alpha'$, which we can write as a fixed point equation $z = f(z)$ for z with $f(z) = 2\alpha' - c - \log(z)$. We can accomplish our goal by determining lower and upper bounds on the fixed point. Since $\frac{d}{dy} f(y) = -1/y < 0$ for $y > 0$, the iteration $z_0 = 2\alpha' - c$ and $z_k = f(z_{k-1})$ is alternating around the fixed point z, provided $z_k > 0$ for all k. Provided $z_0 > 1$, $z_1 = f(z_0) < z_0$, from which we conclude that $z_1 \leq z < z_0$. Since we are assuming that $\alpha \geq 6$ and from above $z_0 \geq 2(\alpha - 1/\alpha) - c$, the condition $z_0 > 1$ is satisfied. If $z_1 \geq 1$, then $0 > \frac{d}{dy} f(y) > -1$ between z_1 and z_0, which implies that z_0 and z_1 are in the region where the iteration converges to z. For our bounds, we only require $z_1 > 0$, so that we can bound z according to $z_1 \leq z \leq z_2$. That

$z_1 > 0$ follows from $\log(y) < y$ for $y > 0$. We have

$$\begin{aligned} z_2 - z_1 &= z_0 - \log(z_1) - (z_0 - \log(z_0)) \\ &= \log(z_0/z_1) \\ &= \log(z_0/(z_0 - \log(z_0))) \\ &= -\log(1 - \log(z_0)/z_0) \\ &= O(\log(z_0)/z_0) = O(\log(\alpha')/\alpha') = O(\log(\alpha)/\alpha), \end{aligned} \tag{2.32}$$

where $z_0 = 2\alpha' - c \in 2\alpha - c + 2[-1, 0]/\alpha$, and so $-\log(z_0) = -\log(2\alpha - c) + O(1/\alpha^2)$. For z_1 we get $z_1 = z_0 - \log(z_0) = 2\alpha - c - \log(2\alpha - c) + O(1/\alpha)$. Applying Eq. 2.32 and from the definitions,

$$q^{-1}(\alpha) = x = \sqrt{2\alpha - c - \log(2\alpha - c) + O(\log(\alpha)/\alpha)}. \tag{2.33}$$

The approximation error in Eq. 2.24 is obtained by expanding the square root. We could have used Newton's method starting from z_0 to obtain better approximations in one step, but the resulting expression is more involved. □

The expression for $\gamma_a(\hat{\theta}; P_{\mathrm{X}})$ confirms our expectation that it approaches the expected value for a standard normal distribution and may be compared to the Berry-Esseen theorem [111]. The empirical endpoint deviation of the CH test approaches that of the exact test for small a (large α). Their squares differ by a term of order $\log(\alpha) = \log|\log(a)|$. Notably, the ratio of the PBR and CH tests' empirical endpoint deviation grows as $\Theta(\sqrt{\log(n)/\alpha})$. The relationships are visualized in Figs. 2.4, 2.5 and 2.6 for different values of a. The figures show that the relative sizes of the empirical endpoint deviations tend toward 1 with smaller a. The $\Theta(\sqrt{\log(n)/\alpha})$ relative growth of the PBR test's endpoint deviations leads to less than a doubling of the deviations relative to the exact test's at $a = 0.01$ and $a = 0.001$ even for $n = 10^6$. So while the test's coverage probabilities are much closer to 1 than the nominal value of $1-a$, we believe that it does not lead to unreasonably conservative results in many applications.

Next we consider the behavior of the true endpoint deviations given by the normalized difference of the true success probability θ and the endpoint obtained from one of the tests. Let

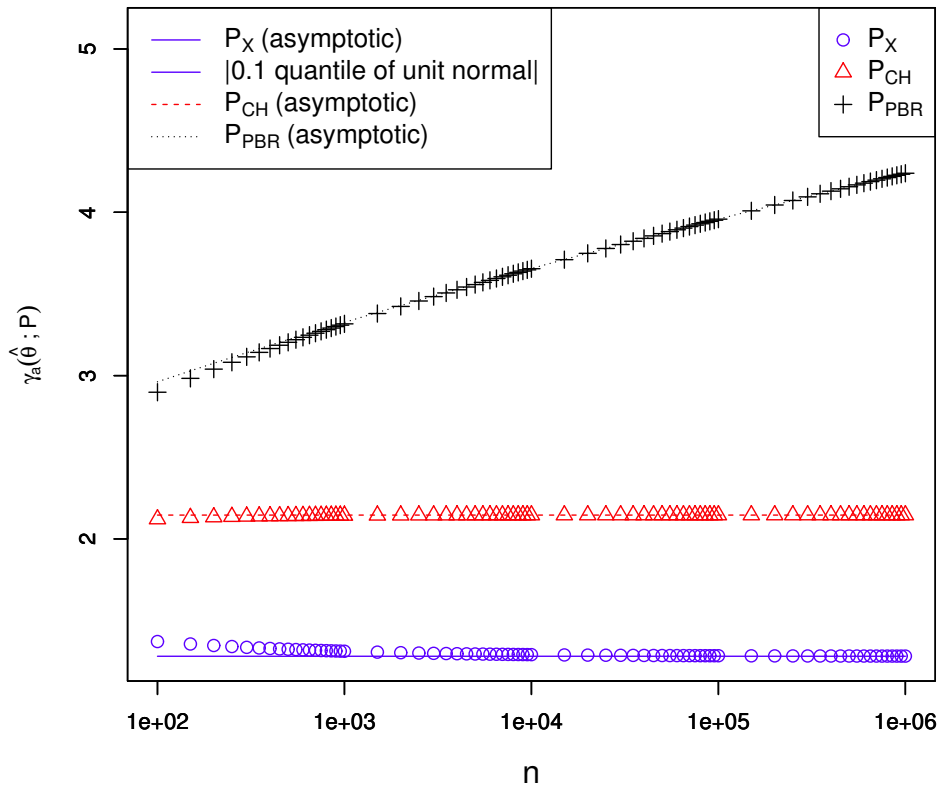

Figure 2.4: Empirical confidence set endpoint deviations at level $a = 0.1$ for $\hat{\theta} = 0.5$ as a function of n. The continuous lines show the expressions obtained after dropping the $O(1/\sqrt{n})$ terms. For the exact test, these expressions are the same as the normal approximation and therefore match the absolute value of the 0.1 quantile of a unit normal.

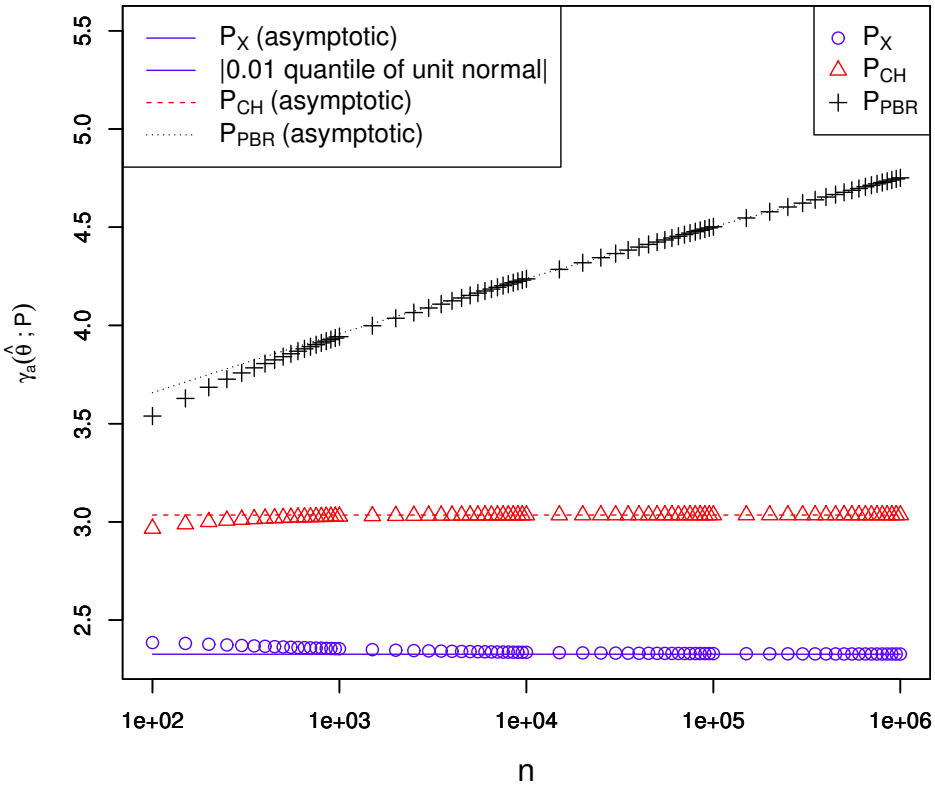

Figure 2.5: Empirical confidence set endpoint deviations at level $a = 0.01$ for $\hat{\theta} = 0.5$ as a function of n. See the caption of Fig. 2.4.

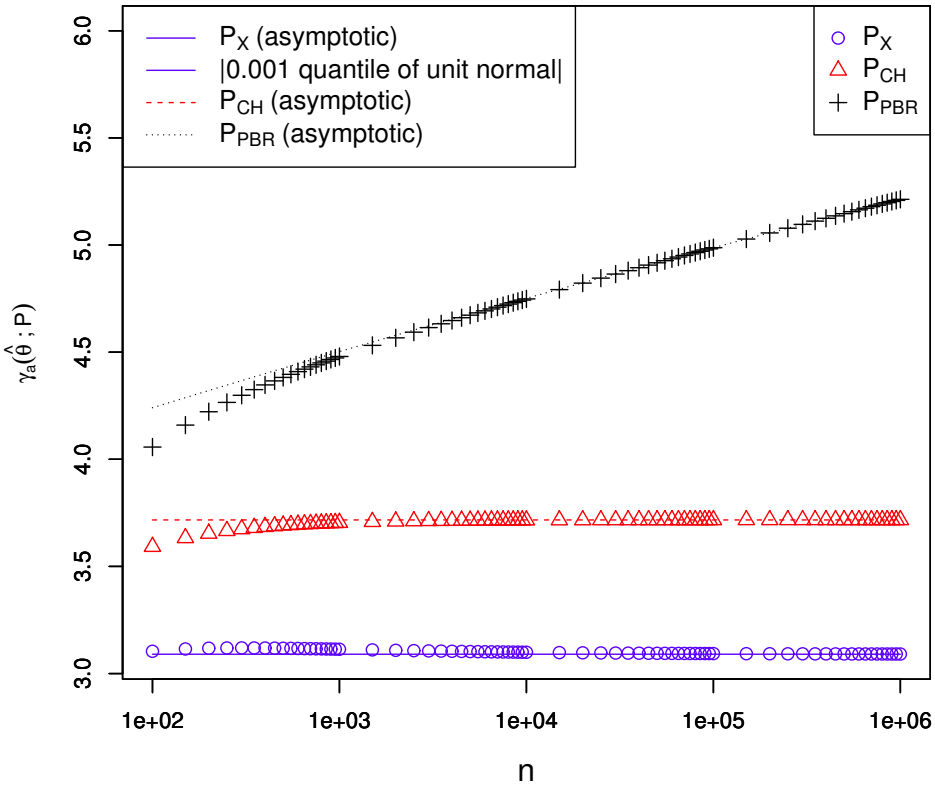

Figure 2.6: Empirical confidence set endpoint deviations at level $a = 0.001$ for $\hat{\theta} = 0.5$ as a function of n. See the caption of Fig. 2.4.

$\sigma = \sqrt{\theta(1-\theta)/n}$ be the true standard deviation and define the true endpoint deviation determined by test P as

$$\begin{aligned}\tilde{\gamma}_a(\hat{\Theta}|P) &= (\theta - \varphi_a(\hat{\Theta}|P))/\sigma \\ &= (\theta - \hat{\Theta})/\sigma + \gamma_a(\hat{\Theta}|P)\hat{\sigma}/\sigma. \end{aligned} \tag{2.34}$$

The true endpoint deviations show how the inferred endpoint compares to θ and therefore directly exhibits the statistical fluctuations of $\hat{\Theta}$. In contrast, the empirical endpoint deviations are to lowest order independent of $\hat{\theta} - \theta$.

We take the view that two tests' endpoints perform similarly if their true endpoint deviations differ by an amount that is small compared to the width of the distribution of the true endpoint deviations. To compare the three tests on this basis, we consider the quantiles for $\hat{\Theta}$ corresponding to $\pm\kappa$ Gaussian standard deviations from θ with κ constant. The quantiles satisfy $\theta_{\pm\kappa} = \theta \pm \kappa\sigma(1 + O(1/\sqrt{n}))$, by the Berry-Esseen theorem or from Thm. 14. Since $\hat{\sigma} = \sigma(1 + O(1/\sqrt{n}))$, we can also see that $\gamma_a(\theta_{\pm\kappa}|P) = \gamma_a(\hat{\theta}|P) + O(1/\sqrt{n})$, and so by substituting into the definition,

$$\tilde{\gamma}_a(\theta_{\pm\kappa}|P) = \gamma_a(\theta|P) \pm \kappa + O(1/\sqrt{n}), \tag{2.35}$$

where the implicit constants depend on κ. For large α, the CH and exact tests' endpoints are close and are dominated by κ, so they perform similarly. But this does not hold for the comparison of the CH or the exact test's endpoints to those of the PBR test, since the latter's endpoint deviation grows as $\sqrt{\log(n)/2}$.

The PBR test's robustness to stopping rules requires that endpoint deviations must grow. Qualitatively, we expect a growth of at least $\Omega(\sqrt{\log\log(n)})$ due to the law of the iterated logarithm. This growth is slower than the $\sqrt{\log(n)/2}$ growth found above, suggesting that improvements are possible, as observed in Ref. [130]. In many applications, the number of trials to be acquired can be determined ahead of time, so full robustness to stopping rules is not necessary. However, the ability to adapt to changing experimental conditions may still be helpful, as the example in Sect. 2.2 shows. If we know the number of trials ahead of time, we can retain the ability to adapt while avoiding the asymptotic growth of the endpoint deviations of the PBR test.

A strategy for avoiding the asymptotic growth of the PBR test's endpoint deviations is to set aside the first $m = \lambda n$ of the trials for training to infer the probability of success, and then use this to determine the test factor to be used on the remaining $(1-\lambda)n$ of the trials. With this strategy, the endpoint deviations are bounded on average and typically. We formalize the training strategy as follows: Modify Eq. 2.11 by setting $F_{k=1} = 1$ for $k < m$ and for $k \geq m$,

$$F_{k+1}(B_{k+1}) = F(B_{k+1}) = \left(\frac{\hat{\Theta}_m}{\varphi}\right)^{B_{k+1}} \left(\frac{1-\hat{\Theta}_m}{1-\varphi}\right)^{1-B_{k+1}}. \tag{2.36}$$

Let $G = F$ if $\varphi \leq \hat{\Theta}_m$ and $G = 1$ otherwise. The G_{k+1} are valid test factors for the null $\mathcal{B}_\varphi$. A p-value for testing $\overline{\mathcal{B}}_\varphi$ is given by

$$P_\lambda(\mathbf{B}|\varphi) = G(1)^{-(n-m)\hat{\Theta}'_m} G(0)^{-(n-m)(1-\hat{\Theta}'_m)} \tag{2.37}$$

where $\hat{\Theta}'_m$ is defined by $(n-m)\hat{\Theta}'_m + m\hat{\Theta}_m = n\hat{\Theta}_n$. We call this the P_λ test.

Define

$$Q_\lambda(\mathbf{B}|\varphi) = \left(\frac{\varphi}{\hat{\Theta}_m}\right)^{(n-m)\hat{\Theta}'_m} \left(\frac{1-\varphi}{1-\hat{\Theta}_m}\right)^{(n-m)(1-\hat{\Theta}'_m)}. \tag{2.38}$$

Then for $\varphi \leq \hat{\Theta}_m$, $Q_\lambda(\mathbf{B}|\varphi) = P_\lambda(\mathbf{B}|\varphi)$. To investigate the behavior of these quantities, we consider values $\mathbf{b}$, $\hat{\theta}$, $\hat{\theta}_m$ and $\hat{\theta}'_m$ of the corresponding RVs. As a function of φ, $Q_\lambda(\mathbf{b}|\varphi)$ is maximized at $\varphi = \hat{\theta}'_m$ and monotone on either side of $\hat{\theta}'_m$. If $\hat{\theta}_m \leq \varphi \leq \hat{\theta}'_m$, then $Q_\lambda(\mathbf{b}|\varphi) \geq 1 = P_\lambda(\mathbf{b}|\varphi)$, So for $\varphi \leq \max(\hat{\theta}_m, \hat{\theta}'_m)$, we can use Q_λ instead of P_λ without changing endpoint calculations.

For determining the endpoint of a level-a one-sided confidence interval from P_λ, we seek the maximum φ such that for all $\varphi' \leq \varphi$, $P_\lambda(\mathbf{b}|\varphi') \leq a$. This maximum value of φ satisfies that $\varphi \leq \min(\hat{\theta}'_m, \hat{\theta}_m)$: For $\hat{\theta}_m \leq \hat{\theta}'_m$, this follows from $P_\lambda(\mathbf{b}|\hat{\theta}_m) = 1$. For $\hat{\theta}_m \geq \hat{\theta}'_m$, the location of the maximum of Q_λ implies that $P_\lambda(\mathbf{b}|\hat{\theta}'_m) \geq P_\lambda(\mathbf{b}|\hat{\theta}_m) = 1$.

We show that endpoint deviations from the P_λ test are typically a constant factor larger than those of the CH test. For large α, the factor approaches $1/\sqrt{1-\lambda}$, approximating the endpoint deviations for a CH test with $(1-\lambda)n$ trials. We begin by comparing P_λ to $P_{\mathrm{CH},(1-\lambda)n}$ with the latter applied to the last $(1-\lambda)n$ trials and $\varphi \leq \hat{\theta}'_m$, where we can use Q_λ in place of P_λ.

$$Q_\lambda(\mathbf{b}|\varphi)/P_{\mathrm{CH},(1-\lambda)n}(\hat{\theta}'_m|\varphi) = \left(\frac{\hat{\theta}'_m}{\hat{\theta}_m}\right)^{(1-\lambda)n\hat{\theta}'_m} \left(\frac{1-\hat{\theta}'_m}{1-\hat{\theta}_m}\right)^{(1-\lambda)n(1-\hat{\theta}'_m)}. \tag{2.39}$$

Or, for the $\log(p)$-value difference l_p,

$$l_p = -\log(Q_\lambda(\mathbf{b}|\varphi)) + \log(P_{\mathrm{CH},(1-\lambda)n}(\hat{\theta}'_m|\varphi)) = -(1-\lambda)n\mathrm{KL}(\hat{\theta}'_m|\hat{\theta}_m), \tag{2.40}$$

which is non-positive. By expanding to second order,

$$\begin{aligned}\mathrm{KL}(t+x|t+y) &= (t+x)\left(\log(1+x/t) - \log(1+y/t)\right) \\ &\quad + (1-t-x)\left(\log(1-x/(1-t)) - \log(1-y/(1-t))\right) \\ &= \frac{(x-y)^2}{2t(1-t)} + O(\max(|x|,|y|)^3).\end{aligned} \tag{2.41}$$

Let $\Delta = \hat{\Theta}_m - \theta$ and $\Delta' = \hat{\Theta}'_m - \theta$. From the above expansion with $t = \theta$, $x = \delta'$ and $y = \delta$ (where δ and δ' are values of Δ and Δ')

$$l_p = -(1-\lambda)n\left(\frac{(\delta-\delta')^2}{2\theta(1-\theta)} + O(\max(|\delta|,|\delta'|^3))\right). \tag{2.42}$$

The RVs Δ and Δ' are independent with means 0 and variances σ^2/λ and $\sigma^2/(1-\lambda)$. Furthermore, $\sqrt{n}\Delta$ and $\sqrt{n}\Delta'$ are asymptotically normal with variances $\theta(1-\theta)/\lambda$ and $\theta(1-\theta)/(1-\lambda)$. Consequently, the RV $\sqrt{n}(\Delta-\Delta')$ is asymptotically normal with variance $v = \theta(1-\theta)/(\lambda(1-\lambda))$. Accordingly, the probability that $n(\Delta-\Delta')^2 \geq \kappa^2\theta(1-\theta)/(\lambda(1-\lambda))$ is asymptotically given by the two-sided tail for κ standard deviations of the standard normal. For determining typical behavior, we consider $(\delta-\delta')^2 = \kappa^2\theta(1-\theta)/(n\lambda(1-\lambda))$ with $\kappa \geq 0$ constant for asymptotic purposes. Observe that $n\Delta^3$ and $n\Delta'^3$ are $\tilde{O}(1/\sqrt{n})$ with probability 1, where the "soft-O" notation $\tilde{O}$ subsumes the polylogarithmic factor from the law of the iterated logarithm. We can now write

$$l_p = -\frac{\kappa^2}{2\lambda} + \tilde{O}(1/\sqrt{n}). \tag{2.43}$$

Fix the level a and thereby also $\alpha = |\log(a)|$. Define $\hat{\sigma}' = \sqrt{\hat{\theta}'_m(1-\hat{\theta}'_m)/(1-\lambda)n}$, and let $\varphi' = \hat{\theta}'_m - \gamma'\hat{\sigma}'$ be the smallest solution of $-\log(Q_\lambda(\mathbf{b}|\varphi')) = \alpha$. Because

$$-\log(Q_\lambda(\mathbf{b}|\varphi')) = -\log(P_{\mathrm{CH},(1-\lambda)n}(\hat{\theta}'_m|\varphi')) + l_p, \tag{2.44}$$

we can estimate γ' as $\gamma' = \gamma_{a',(1-\lambda)n}(\hat{\theta}'_m; P_{\mathrm{CH}}) = \sqrt{2(\alpha-l_p)} + O(\alpha/\sqrt{n})$ with $a' = e^{-(\alpha-l_p)}$. Here, the subscript $(1-\lambda)n$ of $\gamma_{a'}$ makes the previously implicit number of trials explicit.

To finish our comparison, we express the endpoint φ' relative to $\hat{\theta}$. For this, we write

$$\begin{aligned}\varphi' &= \hat{\theta}'_m - \gamma'\hat{\sigma}' \\ &= \hat{\theta} + (\hat{\theta}'_m - \hat{\theta}) - \gamma'\hat{\sigma}\sqrt{\frac{\hat{\theta}'_m(1-\hat{\theta}'_m)}{(1-\lambda)\hat{\theta}(1-\hat{\theta})}} \\ &= \hat{\theta} + (\hat{\theta}'_m - \hat{\theta}) - \frac{\gamma'}{\sqrt{1-\lambda}}\hat{\sigma}\left(1 + O(|\hat{\theta} - \hat{\theta}'_m|)\right). \end{aligned} \tag{2.45}$$

We have $\hat{\theta}'_m - \hat{\theta} = \lambda(\hat{\theta}'_m - \hat{\theta}_m) = \lambda(\delta' - \delta)$, and we are considering the case $\lambda|\delta' - \delta| = \kappa\sqrt{\lambda\theta(1-\theta)/(n(1-\lambda))}$, so

$$\varphi' = \hat{\theta} - \frac{\gamma'}{\sqrt{1-\lambda}}\hat{\sigma}\left(1 + O(1/\sqrt{n})\right). \tag{2.46}$$

We can therefore identify

$$\begin{aligned}\gamma_a(\hat{\theta}|P_\lambda) &= \frac{\gamma'}{\sqrt{1-\lambda}}(1 + O(1/\sqrt{n})) \\ &= \frac{\sqrt{2(\alpha + \kappa^2/(2\lambda) + \tilde{O}(1/\sqrt{n}))} + O(\alpha/\sqrt{n})}{\sqrt{1-\lambda}}(1 + O(1/\sqrt{n})) \\ &= \frac{\sqrt{2(\alpha + \kappa^2/(2\lambda))}}{\sqrt{1-\lambda}} + \tilde{O}(\alpha/\sqrt{n}), \end{aligned} \tag{2.47}$$

which compares as promised to $\gamma_a(\hat{\theta}; P_{\mathrm{CH}}) = \sqrt{2\alpha} + O(\alpha/\sqrt{n})$.

2.6 Conclusion

It is clear that for the specific problem of one-sided hypothesis testing and confidence intervals for Bernoulli RVs, it is always preferable to use the exact test in the ideal case, where the trials are i.i.d. For general nulls, exact tests are typically not available, so approximations are used. The approximations often do not take into account failure of underlying distributional assumptions. The approximation errors can be large at high significance. Thus trustworthy alternatives such as those based on large deviation bounds or test supermartingales are desirable. Our goal here is not to suggest that these alternatives are better for the example of Bernoulli RVs, but to determine the gap between them and an exact test, in a case where an exact test is known and all tests are readily calculable. The suggestion is that for high significance applications, the gaps are relatively small

on the relevant logarithmic scale. For p-values, they are within what is expected from experiment-to-experiment variation, even for moderate significances. For confidence intervals, the increase in size is bounded by a constant if the number of trials is known ahead of time, but there is a slowly growing cost with number of trials if we allow for arbitrary stopping-rules.

2.7 Appendix

2.7.1 Preliminaries

Notation and definitions are as introduced in the text. The p-value bounds obtained by the three tests investigated are denoted by P_{X} for the exact, P_{CH} for the Chernoff-Hoeffding, and P_{PBR} for the PBR test. They depend on n, φ and $\hat{\Theta}$. For reference, here are the definitions again.

$$
\begin{aligned}
P_{\mathrm{X}}(\hat{\Theta}|\varphi,n) &= \sum_{i=n\hat{\Theta}}^{n} \varphi^i(1-\varphi)^{n-i}\binom{n}{i},\\
P_{\mathrm{CH}}(\hat{\Theta}|\varphi) &= \begin{cases} \left(\frac{\varphi}{\hat{\Theta}}\right)^{n\hat{\Theta}} \left(\frac{1-\varphi}{1-\hat{\Theta}}\right)^{n(1-\hat{\Theta})} & \text{if } \hat{\Theta} \geq \varphi,\\ 1 & \text{otherwise.} \end{cases}\\
P_{\mathrm{PBR}}(\hat{\Theta}|\varphi) &= \begin{cases} \varphi^{n\hat{\Theta}}(1-\varphi)^{n(1-\hat{\Theta})}(n+1)\binom{n}{n\hat{\Theta}} & \text{if } \hat{\Theta} \geq \varphi,\\ \hat{\Theta}^{n\hat{\Theta}}(1-\hat{\Theta})^{n(1-\hat{\Theta})}(n+1)\binom{n}{n\hat{\Theta}} & \text{otherwise.} \end{cases}
\end{aligned}
\tag{2.48}
$$

The gain per trial for a p-value bound P_n is $G_n(P_n) = -\log(P_n)/n$. The values of φ, $\hat{\theta}$ and θ are usually constrained. Unless otherwise stated, we assume that $0 < \varphi, \hat{\theta}, \theta < 1$ and $n \geq 1$.

Most of this appendix is dedicated to obtaining upper and lower bounds on $\log(p)$-values and lower bounds on endpoints of confidence intervals. We make sure that the upper and lower bounds differ by quantities that converge to zero as n grows. Their differences are $O(1/n)$ for $\log(p)$-values and $O(1/\sqrt{n})$ for confidence lower bounds. We generally aim for simplicity when expressing these bounds, so we do not obtain tight constants.

2.7.2 Closed-Form Expression for P_{PBR}

Theorem 5. *Define*

$$\begin{aligned}\widetilde{\Theta}_k &= \frac{1}{k+2}\left(S_k+1\right),\\ F_{k+1} &= \left(\frac{\widetilde{\Theta}_k}{\varphi}\right)^{B_{k+1}}\left(\frac{1-\widetilde{\Theta}_k}{1-\varphi}\right)^{1-B_{k+1}}.\end{aligned} \tag{2.49}$$

Then

$$\frac{1}{\prod_{k=1}^{n} F_k} = \varphi^{n\hat{\Theta}}(1-\varphi)^{n(1-\hat{\Theta})}(n+1)\binom{n}{n\hat{\Theta}}. \tag{2.50}$$

Proof. The proof proceeds by induction. Write P_k for the right-hand side of Eq. 2.50. For $n=0$, $P_0=1$, and the left-hand side of Eq. 2.50 evaluates to 1 as required, with the usual convention that the empty product evaluates to 1.

Now suppose that Eq. 2.50 holds at trial $n=k$. For $n=k+1$ we can use $(k+1)\hat{\Theta}_{k+1} = S_{k+1} = S_k + B_{k+1}$. We expand the binomial expression to rewrite the right-hand side as

$$\begin{aligned}P_{k+1} &= \varphi^{k\hat{\Theta}_k+B_{k+1}}(1-\varphi)^{k(1-\hat{\Theta}_k)+(1-B_{k+1})}(k+2)\binom{k+1}{k\hat{\Theta}_k+B_{k+1}}\\ &= \varphi^{k\hat{\Theta}_k}(1-\varphi)^{k(1-\hat{\Theta}_k)}(k+1)\binom{k}{k\hat{\Theta}_k}\\ &\quad\cdot \varphi^{B_{k+1}}(1-\varphi)^{1-B_{k+1}}(k+2)(k-k\hat{\Theta}_k+1)^{-(1-B_{k+1})}(k\hat{\Theta}_k+1)^{-B_{k+1}}.\end{aligned} \tag{2.51}$$

Since $\widetilde{\Theta}_k = (S_k+1)/(k+2) = (k\hat{\Theta}_k+1)/(k+2)$ and $1-\widetilde{\Theta}_k = (k-S_k+1)/(k+2) = (k-k\hat{\Theta}_k+1)/(k+2)$, the identity simplifies to

$$P_{k+1} = P_k \cdot \frac{1}{F_{k+1}}, \tag{2.52}$$

thus establishing the induction step. □

The expression in Eq. 2.50 can be seen as the inverse of a positive martingale for $\mathcal{H}_0 = \{\nu_\varphi\}$ determined by S_n. The complete family of such martingales was obtained by Ville [148], Chapter 5, Sect. 3, Eq. 21. Ours is obtained from Ville's with $dF(t) = dt$ as the probability measure.

2.7.3 Log-p-Value Approximations

We use $-\log(P_{\mathrm{CH},n}(t|\varphi)) = n\mathrm{KL}(t|\varphi)$ as our reference value. According to Thm. 1, the $\log(p)$-values are ordered according to $-\log(P_{\mathrm{PBR}}) \leq -\log(P_{\mathrm{CH}}) \leq -\log(P_{\mathrm{X}})$. To express the asymptotic differences between the $\log(p)$-values, we use auxiliary functions. The first is

$$\begin{aligned} H_n(t) &= -\log\left(t^{nt}(1-t)^{n(1-t)}\binom{n}{nt}\sqrt{n+1}\right) \\ &= -nt\log(t) - n(1-t)\log(1-t) - \log\binom{n}{nt} - \frac{1}{2}\log(n+1). \end{aligned} \tag{2.53}$$

The first two terms of this expression can be recognized as the Shannon entropy of n independent random bits, each with probability t for bit value 1. For $t \in [1/n, 1-1/n]$ and with Stirling's approximation $\sqrt{2\pi n}(n/e)^n e^{1/(12n+1)} \leq n! \leq \sqrt{2\pi n}(n/e)^n e^{1/(12n)}$ applied to the binomial coefficient, we get

$$\begin{aligned} \log\binom{n}{nt} &= \log\left(\frac{n!}{(tn)!((1-t)n)!}\right) \\ &\in \log\left(\frac{\sqrt{2\pi n}}{\sqrt{2\pi tn}\sqrt{2\pi(1-t)n}}\right) + \log\left(\frac{(n/e)^n}{(tn/e)^{tn}((1-t)n/e)^{(1-t)n}}\right) \\ &\quad + \left[\frac{1}{12n+1}, \frac{1}{12n}\right] + \left[-\frac{1}{12tn} - \frac{1}{12(1-t)n}, -\frac{1}{12tn+1} - \frac{1}{12(1-t)n+1}\right] \\ &= -\frac{1}{2}\log(2\pi t(1-t)) - \frac{1}{2}\log(n) - tn\log(t) - (1-t)n\log(1-t) \\ &\quad + \left[\frac{1}{12n+1} - \frac{1}{12t(1-t)n}, \frac{1}{12n} - \frac{12n+2}{(12tn+1)(12(1-t)n+1)}\right]. \end{aligned} \tag{2.54}$$

We can increase the interval to simplify the bounds while preserving convergence for large n. For the lower bound, we use $-1/(12t(1-t)n)$. For the upper bound, note that $(12tn+1)(12(1-t)n+1)$ is maximized at $t = 1/2$. We can therefore increase the upper bound according to

$$\frac{1}{12n} - \frac{12n+2}{(12tn+1)(12(1-t)n+1)} \leq \frac{1}{12n} - \frac{2}{6n+1} \leq 0 \tag{2.55}$$

for $n \geq 1$. From this we obtain the interval expression

$$H_n(t) \in \frac{1}{2}\log(2\pi t(1-t)) - \frac{1}{2}\log(1+1/n) + \left[0, \frac{1}{12nt(1-t)}\right], \tag{2.56}$$

valid for $t \in [1/n, 1-1/n]$. The boundary values of H_n at $t = 0$ and $t = 1$ are $-\log(n+1)/2$.

The next auxiliary function is

$$Y(t) = \frac{1}{e^{-t^2/2}} \int_t^\infty e^{-s^2/2} ds \in \left(\frac{t}{1+t^2}, \frac{1}{t} \right) \quad \text{for } t > 0, \tag{2.57}$$

where the bounds are from Ref. [102]. See this reference for a summary of all properties of Y mentioned here, or Ref. [116] for more details. The function Y is related to the tail of the standard normal distribution, the Q-function, by $Q(t) = e^{-t^2/2} Y(t)/\sqrt{2\pi}$. The function Y is monotonically decreasing, convex, $Y(0) = \sqrt{\pi/2}$, and it satisfies the differential equation $\frac{d}{dt} Y(t) = tY(t) - 1$. We make use of the following bounds involving Y:

$$-\log tY(t) \in \left[0, \frac{1}{t^2} \right]. \tag{2.58}$$

The lower bound comes from the upper bound $1/t$ for $Y(t)$. The upper bound is from the lower bound $t/(1+t^2)$ for $Y(t)$. Specifically, we compute $-\log(Y(t)) \leq -\log(t/(1+t^2)) = \log(t) + \log(1+1/t^2) \leq \log(t) + 1/t^2$.

With these definitions, we can express the $\log(p)$-values in terms of their difference from $-\log(P_{\text{CH}})$.

Theorem 6. *For* $0 < \varphi \leq t < 1$,

$$-\log(P_{\text{PBR},n}(t|\varphi)) = -\log(P_{\text{CH},n}(t|\varphi)) - \frac{1}{2}\log(n+1) + H_n(t) \tag{2.59}$$

$$\in -\log(P_{\text{CH},n}(t|\varphi)) - \frac{1}{2}\log(n+1) + \frac{1}{2}\log(2\pi t(1-t)) - \frac{1}{2}\log(1+1/n) + \left[0, \frac{1}{12nt(1-t)}, \right] \tag{2.60}$$

Proof. The theorem is obtained by substituting definitions and then applying the bounds of Eq. 2.56

on $H_n(t)$. Here are the details.

$$\begin{aligned}
-\log(P_{\mathrm{PBR},n}(t|\varphi)) &= -\log\left(\varphi^{nt}(1-\varphi)^{n(1-t)}(n+1)\binom{n}{nt}\right) \\
&= -\log\left(\left(\frac{\varphi}{t}\right)^{nt}\left(\frac{1-\varphi}{1-t}\right)^{n(1-t)}\right) \\
&\quad -\log\left(t^{nt}(1-t)^{n(1-t))}(n+1)\binom{n}{nt}\right) \\
&= -\log(P_{\mathrm{CH},n}(t|\varphi)) - \frac{1}{2}\log(n+1) \\
&\quad -\log\left(t^{nt}(1-t)^{n(1-t))}\sqrt{n+1}\binom{n}{nt}\right) \\
&= -\log(P_{\mathrm{CH},n}(t|\varphi)) - \frac{1}{2}\log(n+1) + H_n(t). \qquad (2.61)
\end{aligned}$$

It remains to substitute the interval expression for $H_n(t)$. □

Theorem 7. *Define*

$$\mathrm{lE}_n(t|\varphi) = \min\left((t-\varphi)\sqrt{\frac{\pi n}{8\varphi(1-\varphi)}}, 1\right). \qquad (2.62)$$

Then for $0 < \varphi < t < 1$,

$$\begin{aligned}
-\log(P_{\mathrm{X},n}(t|\varphi)) \in &-\log(P_{\mathrm{PBR},n}(t|\varphi)) + \log(n+1) - \log\left(t\sqrt{\frac{(1-\varphi)}{\varphi}}\right) \\
&-\log\left(\sqrt{n}Y\left(\sqrt{\frac{n}{\varphi(1-\varphi)}}(t-\varphi)\right)\right) + \left[-\frac{\mathrm{lE}_n(t|\varphi)}{n(t-\varphi)}, 0\right], \qquad (2.63) \\
-\log(P_{\mathrm{X},n}(t|\varphi)) \in &-\log(P_{\mathrm{CH},n}(t|\varphi)) + \frac{1}{2}\log(n) - \log\left(\sqrt{\frac{t(1-\varphi)}{2\pi(1-t)\varphi}}\right) \\
&-\log\left(\sqrt{n}Y\left(\sqrt{\frac{n}{\varphi(1-\varphi)}}(t-\varphi)\right)\right) + \left[-\frac{\mathrm{lE}_n(t|\varphi)}{n(t-\varphi)}, \frac{1}{12nt(1-t)}\right]. \qquad (2.64)
\end{aligned}$$

Observe that $\mathrm{lE}_n(t|\varphi)$ is $O(1)$ with respect to n for $t > \varphi$ constant. The first term in the defining minimum is smaller than 1 only for φ within less than one standard deviation (which is $O(1/\sqrt{n})$) of t. It is defined so that the primary dependence on the parameters is visible in the interval bounds.

Proof. For approximating P_{X}, we apply Thm. 2 of Ref. [102] with the following sequence of sub-

stitutions, the first four of which expand the definitions in the reference:

$$\begin{aligned} B(k;n,p) &\leftarrow \sum_{j=k}^{n} b(j;n,p), \\ b(k-1;n-1,p) &\leftarrow \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k}, \\ x &\leftarrow (k-pn)/\sigma, \\ \sigma &\leftarrow \sqrt{np(1-p)}, \\ p &\leftarrow \varphi, \\ k &\leftarrow nt. \end{aligned} \tag{2.65}$$

With the given substitutions and $Y(t)$ as defined by Eq. 2.57, we obtain for $t \geq \varphi$,

$$\begin{aligned} -\log(P_{\mathrm{X}}) \in & -\log\left(\sqrt{n\varphi(1-\varphi)}\varphi^{nt-1}(1-\varphi)^{n(1-t)}\binom{n-1}{nt-1}\right) \\ & -\log\left(Y\left(\frac{\sqrt{n}(t-\varphi)}{\sqrt{\varphi(1-\varphi)}}\right)\right) + \left[-\frac{\mathrm{lE}_n(t|\varphi)}{n(t-\varphi)}, 0\right] \\ = & -\log\left(\varphi^{nt}(1-\varphi)^{n(1-t)}(n+1)\binom{n}{nt}\right) - \log\left(\frac{nt\sqrt{n\varphi(1-\varphi)}}{\varphi n(n+1)}\right) \\ & -\log\left(Y\left(\frac{\sqrt{n}(t-\varphi)}{\sqrt{\varphi(1-\varphi)}}\right)\right) + \left[-\frac{\mathrm{lE}_n(t|\varphi)}{n(t-\varphi)}, 0\right] \\ = & -\log(P_{\mathrm{PBR}}) + \log(n+1) - \log\left(t\sqrt{\frac{(1-\varphi)}{\varphi}}\right) \\ & -\log\left(\sqrt{n}Y\left(\frac{\sqrt{n}(t-\varphi)}{\sqrt{\varphi(1-\varphi)}}\right)\right) + \left[-\frac{\mathrm{lE}_n(t|\varphi)}{n(t-\varphi)}, 0\right]. \end{aligned} \tag{2.66}$$

The second identity of the theorem follows by substituting the expression from Thm. 6. □

We can eliminate the function Y from the expressions by applying the bounds from Eq. 2.58.

Corollary 8. *With the assumptions of Thm. 7,*

$$\begin{aligned} -\log(P_{\mathrm{X},n}(t|\varphi)) \in & -\log(P_{\mathrm{CH},n}(t|\varphi)) + \frac{1}{2}\log(n) - \log\left(\frac{1-\varphi}{t-\varphi}\sqrt{\frac{t}{2\pi(1-t)}}\right) \\ & + \left[-\frac{\mathrm{lE}_n(t|\varphi)}{n(t-\varphi)}, \frac{\varphi(1-\varphi)}{(t-\varphi)^2 n} + \frac{1}{12nt(1-t)}\right]. \end{aligned} \tag{2.67}$$

Proof. Define $c = (t-\varphi)/\sqrt{\varphi(1-\varphi)}$. In view of Eq. 2.58, we have

$$\begin{aligned} -\log\left(\sqrt{n}Y\left(\sqrt{\frac{n}{\varphi(1-\varphi)}}(t-\varphi)\right)\right) &= -\log(\sqrt{n}Y(c\sqrt{n})) \\ &= \log(c) - \log(c\sqrt{n}Y(c\sqrt{n})) \\ &\in \log(c) + \left[0, \frac{1}{c^2 n}\right]. \end{aligned} \tag{2.68}$$

Substituting in Eq. 2.64 and simplifying the expression gives the desired result. □

2.7.4 Asymptotic Normality of the $\log(p)$-Values and Their Differences

The main tool for establishing the asymptotic distribution of the $\log(p)$-values is the "delta method". A version sufficient for our purposes is Thm. 1.12 and Cor. 1.1 of Ref. [133]. The notation $X_n \xrightarrow{D} N(\mu, \sigma^2)$ means that X_n converges in distribution to the normal distribution with mean μ and variance σ^2. By the central limit theorem, $\hat{\Theta}_n = S_n/n$ satisfies $\sqrt{n}(\hat{\Theta}_n - \theta) \xrightarrow{D} N(0, \theta(1-\theta))$. An application of the delta method therefore yields the next lemma.

Lemma 9. *Let $F : \mathbb{R} \to \mathbb{R}$ be differentiable at θ, with $F'(\theta) \neq 0$. Then*

$$\sqrt{n}(F(\hat{\Theta}_n) - F(\theta)) \xrightarrow{D} N\left(0, F'(\theta)^2\theta(1-\theta)\right)$$

Theorem 10. *For $P = P_{\mathrm{CH}}$, $P = P_{\mathrm{PBR}}$ or $P = P_{\mathrm{X}}$, and $0 < \varphi < \theta < 1$ constant, the gain per trial $G_n(P)$ converges in distribution according to*

$$\sqrt{n}(G_n(P) - \mathrm{KL}(\theta|\varphi)) \xrightarrow{D} N(0, \sigma_G^2), \tag{2.69}$$

with

$$\sigma_G^2 = \theta(1-\theta)\left(\log\left(\frac{\theta}{1-\theta}\frac{1-\varphi}{\varphi}\right)\right)^2.$$

Proof. Consider $P = P_{\mathrm{CH}}$ first. In Lem. 9, define $F(x) = \mathrm{KL}(x|\varphi) = x\log(x/\varphi) + (1-x)\log((1-x)/(1-\varphi))$ so that $F(\hat{\Theta}_n) = G_n(P_{\mathrm{CH}})$. For the derivative of F at $x = \theta$, we get

$$F'(\theta) = \log\left(\frac{\theta}{1-\theta}\frac{1-\varphi}{\varphi}\right). \tag{2.70}$$

The theorem now follows for P_{CH} by applying Lem. 9.

Thm. 6 and the law of large numbers imply that $(-\log(P_{\mathrm{PBR}})/\sqrt{n}) - (-\log(P_{\mathrm{CH}})/\sqrt{n})$ converges in probability to 0. Cor. 8 implies the same for P_{X}, namely that $(-\log(P_X)/\sqrt{n}) - (-\log(P_{\mathrm{CH}})/\sqrt{n})$ converges in probability to 0. In general, if $X_n - Y_n$ converges in probability to 0 and $Y_n \xrightarrow{D} \mu$, then $X_n \xrightarrow{D} \mu$, see Ref. [25], Prop. 6.3.3. The statement of the theorem to be proven now follows for $P = P_{\mathrm{PBR}}$ and $P = P_{\mathrm{X}}$ by comparison of $\sqrt{n}G_n(P_{\mathrm{PBR}})$ and $\sqrt{n}G_n(P_{\mathrm{X}})$ to $\sqrt{n}G_n(P_{\mathrm{CH}})$. □

The differences of the $\log(p)$-values have much tighter distributions. They are also asymptotically normal with scaling and variances given in the next theorem. The differences are $\Omega(\log(n))$ with standard deviations $O(1/\sqrt{n})$.

Theorem 11. *Let $0 < \varphi < \theta < 1$ be constant. If $\theta \neq 1/2$, then $P_{\mathrm{PBR}}/(\sqrt{n}P_{\mathrm{CH}})$ satisfies*

$$-\sqrt{n}\log\left(\frac{\sqrt{2\pi\theta(1-\theta)}P_{\mathrm{PBR}}}{\sqrt{n}P_{\mathrm{CH}}}\right) \xrightarrow{D} N\left(0, \frac{(1-2\theta)^2}{4\theta(1-\theta)}\right). \tag{2.71}$$

If $\varphi \neq \theta(2\theta-1)$, then $\sqrt{n}P_{\mathrm{X}}/P_{\mathrm{CH}}$ satisfies

$$-\sqrt{n}\log\left(\frac{\theta-\varphi}{1-\varphi}\sqrt{\frac{2\pi(1-\theta)}{\theta}}\frac{\sqrt{n}P_{\mathrm{X}}}{P_{\mathrm{CH}}}\right) \xrightarrow{D} N\left(0, \frac{(\theta(1-2\theta)+\varphi)^2}{4(\theta-\varphi)^2\theta(1-\theta)}\right), \tag{2.72}$$

Proof. From Thm. 6, Eq. 2.60 and the law of large numbers, we see that

$$\sqrt{n}\left(-\log\left(\frac{P_{\mathrm{PBR}}}{\sqrt{n}P_{\mathrm{CH}}}\right) - \log\left(\sqrt{2\pi\hat{\Theta}(1-\hat{\Theta})}\right)\right) \tag{2.73}$$

converges in probability to zero. From Lem. 9 and

$$\frac{d}{dx}\log(x(1-x))/2 = \frac{1}{2x} - \frac{1}{2(1-x)} = \frac{1-2x}{2x(1-x)}, \tag{2.74}$$

we conclude

$$\sqrt{n}\left(\log\left(\sqrt{2\pi\hat{\Theta}(1-\hat{\Theta})}\right) - \log\left(\sqrt{2\pi\theta(1-\theta)}\right)\right) \xrightarrow{D} N\left(0, \left(\frac{1-2\theta}{2\theta(1-\theta)}\right)^2\theta(1-\theta)\right). \tag{2.75}$$

Combining the above observations gives Eq. 2.71.

Similarly, from Cor. 8 and taking note of the definition of $\mathrm{lE}_n(t|\varphi)$,

$$\sqrt{n}\left(-\log\left(\frac{\sqrt{n}P_{\mathrm{X}}}{P_{\mathrm{CH}}}\right) - \log\left(\frac{\hat{\Theta}-\varphi}{1-\varphi}\sqrt{\frac{2\pi(1-\hat{\Theta})}{\hat{\Theta}}}\right)\right) \tag{2.76}$$

converges in probability to zero. The relevant derivative is

$$\frac{d}{dx}\left(\log(x-\varphi)+\log((1-x)/x)/2\right)=\frac{1}{x-\varphi}-\frac{1}{2(1-x)}-\frac{1}{2x}=\frac{x(1-2x)+\varphi}{2(x-\varphi)x(1-x)}, \tag{2.77}$$

from which

$$\sqrt{n}\left(\log\left(\frac{\hat{\Theta}-\varphi}{1-\varphi}\sqrt{\frac{2\pi(1-\hat{\Theta})}{\hat{\Theta}}}\right)-\log\left(\frac{\theta-\varphi}{1-\varphi}\sqrt{\frac{2\pi(1-\hat{\theta})}{\hat{\theta}}}\right)\right) \xrightarrow{D} N\left(0,\left(\frac{\theta(1-2\theta)+\varphi}{2(\theta-\varphi)\theta(1-\theta)}\right)^2\theta(1-\theta)\right), \tag{2.78}$$

and combining the two observations gives Eq. 2.72. □

2.7.5 Confidence Interval Endpoints

For the one-sided confidence intervals, we need to determine the lower boundaries of acceptance regions, that is the confidence lower bounds. By monotonicity of the p-values in φ, it suffices to solve equations of the form $-\log(P(\hat{\theta},\varphi))=\alpha$, where $a=e^{-\alpha}$ is the desired significance level. Here we obtain lower and upper bounds on the solutions φ.

To illuminate the asymptotic behavior of solutions φ of $-\log(P(\hat{\theta},\varphi))=\alpha$, we reparametrize the log-p-values so that our scale is set by an empirical standard deviation, namely $\hat{\sigma}=\sqrt{\hat{\theta}(1-\hat{\theta})/n}$. Thus we express the solution as

$$\varphi(\gamma,\hat{\theta})=\hat{\theta}-\hat{\sigma}\gamma, \tag{2.79}$$

in terms of a scaled deviation down from $\hat{\theta}$. Inverting for γ we get

$$\gamma=\gamma(\varphi,\hat{\theta})=\frac{\hat{\theta}-\varphi}{\hat{\sigma}}. \tag{2.80}$$

Theorem 12. *Let $0<\hat{\theta}<1$ and $\alpha>0$. Suppose that $\alpha\le n\hat{\theta}^2(1-\hat{\theta})^2/8$. Then there is a solution $\gamma_\alpha>0$ of the identity $-\log(P_{\mathrm{CH}}(\hat{\theta},\varphi(\gamma_\alpha,\hat{\theta})))=\alpha$ satisfying*

$$\gamma_\alpha\in\sqrt{2\alpha}\left(1+\frac{5}{2}\frac{\sqrt{\alpha}}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1]\right)^{-1/2}. \tag{2.81}$$

The constants in this theorem and elsewhere are chosen for convenience, not for optimality; better constants can be extracted from the proofs. Note that the upper bound on α ensures that

the reciprocal square root is bounded away from zero. However, for the relative error to go to zero as n grows requires $\alpha = o(n)$.

Proof. Consider the parametrized bound $\alpha \leq 2n\hat{\theta}^2(1-\hat{\theta})^2(1-a_1)^2$, where later we set $a_1 = 3/4$ to match the theorem statement. Let $F(\gamma) = -\log(P_{\mathrm{CH}}(\hat{\theta}, \varphi(\gamma, \hat{\theta})))$. F is continuous and monotone increasing. A standard simplification of the Chernoff-Hoeffding bound noted in Ref. [69] is

$$P_{\mathrm{CH}} \leq e^{-2n(\hat{\theta}-\varphi)^2} = e^{-2\hat{\theta}(1-\hat{\theta})\gamma^2}. \tag{2.82}$$

For $\varphi = \varphi(\gamma_\alpha, \hat{\theta})$ solving the desired equation, we have $(\hat{\theta}-\varphi) \leq \sqrt{\alpha/(2n)}$ (by monotonicity), which in turn is bounded above according to $\sqrt{\alpha/2n} \leq \hat{\theta}(1-\hat{\theta})(1-a_1) \leq \hat{\theta}(1-a_1)$, according to our assumed bound. We conclude that $\varphi \geq a_1\hat{\theta}$. For the solution γ_α, we get $\gamma_\alpha \leq \sqrt{\alpha/(2\hat{\theta}(1-\hat{\theta}))} \leq \sqrt{n\hat{\theta}(1-\hat{\theta})}(1-a_1)$.

We now Taylor expand $\mathrm{KL}(\hat{\theta}|\varphi)$ with remainder at third order around $\varphi = \hat{\theta}$. Write $f(x) = \mathrm{KL}(\hat{\theta}|\hat{\theta}-x)$, where we can restrict x according to $\hat{\theta} \geq \hat{\theta} - x = \varphi \geq a_1\hat{\theta}$. The derivatives of f can be written explicitly as follows:

$$f^{(k)}(x) = (k-1)!\frac{\hat{\theta}}{(\hat{\theta}-x)^k} - (-1)^{k-1}(k-1)!\frac{1-\hat{\theta}}{(1-\hat{\theta}+x)^k}. \tag{2.83}$$

We have

$$\begin{aligned}
f^{(1)}(0) &= 0, \\
f^{(2)}(0) &= \frac{1}{\hat{\theta}} + \frac{1}{1-\hat{\theta}} = \frac{1}{\hat{\theta}(1-\hat{\theta})}, \\
f^{(3)}(x) &= 2\frac{\hat{\theta}}{(\hat{\theta}-x)^3} - 2\frac{1-\hat{\theta}}{(1-\hat{\theta}+x)^3}, \\
f^{(3)}(x) &\leq 2\frac{\hat{\theta}}{(\hat{\theta}-x)^3} \leq 2\frac{\hat{\theta}}{a_1^3\hat{\theta}^3} = 2\frac{1}{a_1^3\hat{\theta}^2}, \\
f^{(3)}(x) &\geq -2\frac{1-\hat{\theta}}{(1-\hat{\theta}+x)^3} \geq -2\frac{1-\hat{\theta}}{(1-\hat{\theta})^3} = -2\frac{1}{(1-\hat{\theta})^2},
\end{aligned} \tag{2.84}$$

since $0 < a_1 < 1$. We use the bounds on $f^{(3)}(x)$ to bound the remainder in the Taylor expansion, where, to get cleaner expressions, we can decrease $\hat{\theta}$ and $1-\hat{\theta}$ to $\hat{\theta}(1-\hat{\theta})$ in the denominators.

$$\mathrm{KL}(\hat{\theta}|\hat{\theta}-x) \in \frac{x^2}{2\hat{\theta}(1-\hat{\theta})} + \frac{x^3}{3(\hat{\theta}(1-\hat{\theta}))^2}\left[-1, \frac{1}{a_1^3}\right]. \tag{2.85}$$

Substituting $x = \gamma_\alpha \sqrt{\hat{\theta}(1-\hat{\theta})/n}$ gives

$$\alpha = -\log(P_{\mathrm{CH}}(\hat{\theta}, \varphi(\gamma_\alpha, \hat{\theta}))) = n\mathrm{KL}(\hat{\theta}|\hat{\theta} - x) \in \frac{\gamma_\alpha^2}{2}\left(1 + \frac{2\gamma_\alpha}{3\sqrt{n\hat{\theta}(1-\hat{\theta})}}\left[-1, \frac{1}{a_1^3}\right]\right). \tag{2.86}$$

For $\hat{\theta} \leq 1/2$, $f^{(4)}(x)$ and $f^{(3)}(0)$ are non-negative, so we could have taken the lower bound in the interval to be zero for $\theta \leq 1/2$. For the theorem, we prefer not to separate the cases.

We substitute the bound $\gamma \leq \sqrt{n\hat{\theta}(1-\hat{\theta})}(1-a_1)$ for the γ multiplying the interval in Eq. 2.86 and use the lower bound in the interval for the inequality

$$\alpha \geq \frac{\gamma^2}{2}\left(1 - \frac{2(1-a_1)}{3}\right). \tag{2.87}$$

For the theorem, we have $a_1 = 3/4$, so $1 - 2(1-a_1)/3 = 5/6$. Inverting the inequality for γ gives $\gamma \leq 2\sqrt{3/5}\sqrt{\alpha}$. Now substituting this bound on γ for the γ multiplying the interval in Eq. 2.86 gives

$$\alpha \in \frac{\gamma^2}{2}\left(1 + \frac{4}{\sqrt{15}}\frac{\sqrt{\alpha}}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}\left[-1, \frac{4^3}{3^3}\right]\right). \tag{2.88}$$

By monotonicity of the appropriate operations,

$$\gamma \in \sqrt{2\alpha}\left(1 + \frac{4}{\sqrt{15}}\frac{\sqrt{\alpha}}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}\left[-1, \frac{4^3}{3^3}\right]\right)^{-1/2}. \tag{2.89}$$

For the theorem statement, we simplify the bounds with $1 \leq 4^3/3^3$ and $4^4/(3^3\sqrt{15}) \leq 5/2$. □

Theorem 13. *Let $0 < \hat{\theta} < 1$ and $\alpha > 0$. Define $\Delta = \log(n+1)/2 - H_n(\hat{\theta})$. Suppose that $\alpha + \Delta \leq n\hat{\theta}^2(1-\hat{\theta})^2/8$. Then there is a solution $\gamma_\alpha > 0$ of the identity $-\log(P_{\mathrm{PBR}}(\hat{\theta}, \varphi(\gamma_\alpha, \hat{\theta}))) = \alpha$ satisfying*

$$\gamma_\alpha \in \sqrt{2(\alpha+\Delta)}\left(1 + \frac{5}{2}\frac{\sqrt{\alpha+\Delta}}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1, 1]\right)^{-1/2}.$$

Proof. By Thm. 6, $-\log(P_{\mathrm{CH}}) - (-\log(P_{\mathrm{PBR}})) = \Delta$. If we define $\tilde{\alpha} = \alpha + \Delta$, then solving $-\log(P_{\mathrm{PBR}}) = \alpha$ is equivalent to solving $-\log(P_{\mathrm{CH}}) = \tilde{\alpha}$. Since Δ depends only on n and $\hat{\theta}$, $\tilde{\alpha}$ does not depend on γ. We can therefore apply Thm. 12 to get the desired bounds. □

Theorem 14. *For $x \geq 0$, let $q(x) = -\log(e^{-x^2/2}Y(x)/\sqrt{2\pi}) = x^2/2 + \log(2\pi)/2 - \log(Y(x))$. Suppose that $0 < \hat{\theta} < 1$, and $\log(2) < \alpha \leq n\hat{\theta}^2(1-\hat{\theta})^2/8$. Then there is a solution γ_α of the identity $-\log(P_{\mathrm{X}}(\hat{\theta}, \varphi(\gamma_\alpha, \hat{\theta})) = \alpha$ satisfying*

$$\gamma_\alpha \in \max\left(0, q^{-1}\left(\alpha\left(1 + \frac{64\sqrt{\alpha}/(15\sqrt{15})}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1]\right) + \frac{\sqrt{\pi/6} + 8\sqrt{\alpha}/\sqrt{15}}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1]\right)\right) \times \left(1 + \frac{2\sqrt{\alpha}/\sqrt{5}}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1]\right), \tag{2.90}$$

where we extend q^{-1} to negative values by $q^{-1}(y) = -\infty$ for $y \leq 0$ (if necessary) when evaluating this interval expression.

The function $q(x)$ is the negative logarithm of the Q-function, which is the tail of the standard normal distribution. The lower bound on α in Thm. 14 ensures that there is a solution with $\gamma_\alpha > 0$, because $q(0) = \log(2)$. For reference, the constants multiplying the interval expressions are $64/(15\sqrt{15}) \approx 1.102$, $8/\sqrt{15} \approx 2.066$, $\sqrt{\pi/6} \approx 0.724$, $2/\sqrt{5} \approx 0.894$. Note that in the large n limit, where the $O(1/\sqrt{n})$ terms are negligible, the value of γ_α in Thm. 14 corresponds to the $(1 - e^{-\alpha})$-quantile of the standard normal.

By monotonicity of q^{-1}, the explicit bounds in Eq. 2.90 are obtained by combining the lower or the upper bounds in intervals in the expression. We remark that q^{-1} behaves well with respect to relative error for α large enough because of the inequalities

$$\begin{aligned} q^{-1}(y)/(1+q^{-1}(y)^2) \leq \frac{d}{dy}q^{-1}(y) \leq 1/q^{-1}(y), & \\ q^{-1}(y)^2 \geq y - q(1) + 1, & \quad \text{for } y \geq q(1) \approx 1.841, \\ q^{-1}(y)^2 \leq 2(y - \log(2)), & \quad \text{for } y \geq q(0) = \log(2), \end{aligned} \tag{2.91}$$

which we now establish. By implicit differentiation and from the properties of Y noted after Eq. 2.57, $\frac{d}{dy}q^{-1}(y)|_{y=q(x)} = Y(x) \in [x/(1+x^2), 1/x]$. Therefore $q^{-1}(y)/(1+q^{-1}(y)^2) \leq \frac{d}{dy}q^{-1}(y) \leq 1/q^{-1}(y)$. For $y \geq \log(2)$, we can integrate $\frac{d}{dz}q^{-1}(z)^2 = 2q^{-1}(z)\frac{d}{dz}q^{-1}(z) \leq 2$ from $z = \log(2)$ to y to show that $q^{-1}(y)^2 = q^{-1}(y)^2 - q^{-1}(\log(2))^2 \leq 2(y - \log(2))$, making use of the identity

$q^{-1}(\log(2)) = 0$. Consider $y, z \geq q(1)$. Since $q^{-1}(z)$ and $0 \leq x \mapsto x^2/(1+x^2)$ are monotone increasing, $q^{-1}(z)^2/(1+q^{-1}(z)^2) \geq q^{-1}(q(1))^2/(1+q^{-1}(q(1))^2) = 1/2$, so the integral of $\frac{d}{dz}q^{-1}(z)^2$ from $z = q(1)$ to y with the lower bound on $\frac{d}{dz}q^{-1}(z)$ gives $q^{-1}(y)^2 - q^{-1}(q(1))^2 = q^{-1}(y)^2 - 1 \geq y - q(1)$.

From the inequality $\frac{d}{dy}q^{-1}(y) \leq 1/q^{-1}(y)$ in Eq. 2.91, integration and monotonicity, for $0 \leq z \leq \delta$,

$$\begin{aligned} q^{-1}(\alpha - z) &\geq q^{-1}(\alpha) - \frac{z}{q^{-1}(\alpha - \delta)} \leq q^{-1}(\alpha)\left(1 - \frac{z}{q^{-1}(\alpha-\delta)^2}\right), \\ q^{-1}(\alpha + z) &\leq q^{-1}(\alpha) + \frac{z}{q^{-1}(\alpha - \delta)} \geq q^{-1}(\alpha)\left(1 + \frac{z}{q^{-1}(\alpha-\delta)^2}\right). \end{aligned} \tag{2.92}$$

To determine the relative error, write $\delta' = \delta/\alpha$ to obtain the interval inclusion

$$q^{-1}(\alpha(1+\delta'[-1,1])) \subseteq q^{-1}(\alpha)\left(1 + \frac{\alpha\delta'}{q^{-1}(\alpha(1-\delta'))^2}[-1,1]\right). \tag{2.93}$$

For $\alpha(1-\delta') > q(1)$, the interval relationship can be weakened to

$$q^{-1}(\alpha(1+\delta'[-1,1])) \subseteq q^{-1}(\alpha)\left(1 + \frac{\alpha\delta'}{\alpha(1-\delta') - q(1) + 1}[-1,1]\right). \tag{2.94}$$

The relative error on the right-hand side is given by the term multiplying the interval, and can be written as $\alpha\delta'/(\alpha - (\alpha\delta' + q(1) - 1))$. If $\alpha\delta' + q(1) - 1 \leq \alpha/2$, then the relative error is bounded by $2\delta'$ which is twice the relative error of α. Of course, for the interval bounds to converge, we need $\alpha = o(n)$.

Proof. As in the proof of Thm. 12, consider the parametrized bound $\alpha \leq 2n\hat{\theta}^2(1-\hat{\theta})^2(1-a_1)^2$, where later we set $a_1 = 3/4$ to match the statement of Thm. 14. From the Chernoff-Hoeffding bound, we get $\varphi \geq a_1\hat{\theta}$ and $\gamma_\alpha \leq \sqrt{\alpha/(2\hat{\theta}(1-\hat{\theta}))} \leq \sqrt{n\hat{\theta}(1-\hat{\theta})}(1-a_1)$.

Define $\tilde{\gamma} = (\hat{\theta} - \varphi)/\sqrt{\varphi(1-\varphi)/n}$. We start from Eq. 2.64, rewritten as follows:

$$\begin{aligned} -\log(P_{\mathrm{X}}) \in n\mathrm{KL}(\hat{\theta}|\varphi) + \frac{1}{2}\log(2\pi) - \log Y(\tilde{\gamma}) - \frac{1}{2}\log\left(\frac{\hat{\theta}(1-\varphi)}{(1-\hat{\theta})\varphi}\right) \\ + \left[-\frac{\mathrm{lE}_n(\hat{\theta}|\varphi)}{n(\hat{\theta}-\varphi)}, \frac{1}{12n\hat{\theta}(1-\hat{\theta})}\right]. \end{aligned} \tag{2.95}$$

If $\tilde{\gamma} \geq \sqrt{8/\pi} \approx 1.6$, $\mathrm{lE}_n(\hat{\theta}|\varphi) = 1$. For better bounds at small values of $\tilde{\gamma}$, we use the other alternative in the definition of lE_n, according to which the lower bound in the last interval of Eq. 2.95 is

$$-\frac{\mathrm{lE}_n(\hat{\theta}|\varphi)}{n(\hat{\theta}-\varphi)} \geq -\frac{\sqrt{\pi/8}}{\sqrt{n\varphi(1-\varphi)}} \geq -\frac{\sqrt{\pi/8}}{\sqrt{na_1\hat{\theta}(1-\varphi)}} \geq -\frac{\sqrt{\pi/8}}{\sqrt{na_1\hat{\theta}(1-\hat{\theta})}}. \tag{2.96}$$

Next we approximate $n\mathrm{KL}(\hat{\theta}|\varphi)$ in terms of $\tilde{\gamma}$ instead of γ. We still write the interval bounds in terms of γ. Let $f(x) = \mathrm{KL}(\varphi + x|\varphi)$. We are concerned with the range $0 \leq x \leq \hat{\theta} - \varphi$, with $\varphi \geq a_1\hat{\theta}$. We have

$$\begin{aligned}
f^{(1)}(x) &= \log((\varphi+x)/\varphi) - \log((1-\varphi-x)/(1-\varphi)) \\
f^{(2)}(x) &= \frac{1}{\varphi+x} + \frac{1}{1-\varphi-x} \\
&= \frac{1}{(\varphi+x)(1-\varphi-x)} \\
f^{(3)}(x) &= -\frac{1}{(\varphi+x)^2} + \frac{1}{(1-\varphi-x)^2} \\
&= -\frac{1-2(\varphi+x)}{(\varphi+x)^2(1-\varphi-x)^2} \\
|f^{(3)}(x)| &\leq \frac{1}{a_1^2\hat{\theta}^2(1-\hat{\theta})^2},
\end{aligned} \tag{2.97}$$

yielding

$$\mathrm{KL}(\varphi + x|\varphi) \in \frac{x^2}{2\varphi(1-\varphi)} + \frac{x^3}{6a_1^2\hat{\theta}^2(1-\hat{\theta})^2}[-1,1], \tag{2.98}$$

and with $x = \tilde{\gamma}\sqrt{\varphi(1-\varphi)/n} = \gamma\sqrt{\hat{\theta}(1-\hat{\theta})/n}$,

$$n\mathrm{KL}(\hat{\theta}|\varphi) \in \frac{\tilde{\gamma}^2}{2} + \frac{\gamma^3}{6a_1^2\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1]. \tag{2.99}$$

For the fourth term on the right-hand side of Eq. 2.95,

$$\frac{d}{dx}\log\left(\frac{\hat{\theta}(1-\hat{\theta}+x)}{(1-\hat{\theta})(\hat{\theta}-x)}\right) = \frac{1}{1-\hat{\theta}+x} + \frac{1}{\hat{\theta}-x} = \frac{1}{(1-\hat{\theta}+x)(\hat{\theta}-x)}, \tag{2.100}$$

whose absolute value is bounded by $1/(a_1\hat{\theta}(1-\hat{\theta}))$ for x in the given range. Thus

$$\log\left(\frac{\hat{\theta}(1-\varphi)}{(1-\hat{\theta})\varphi}\right) \in \frac{\gamma}{a_1\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1]. \tag{2.101}$$

Since $P_{\mathrm{X}} \leq P_{\mathrm{CH}}$, we can also use the bound $\gamma \leq 2\sqrt{3/5}\sqrt{\alpha}$ obtained in the proof of Thm. 12. Substituting $a_1 = 3/4$ as needed, the equation to solve is now

$$\begin{aligned}\alpha \in \frac{\tilde{\gamma}^2}{2} + \frac{1}{2}\log(2\pi) - \log Y(\tilde{\gamma}) \\ + \frac{8}{\sqrt{15}}\frac{\sqrt{\alpha}}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1] + \frac{64}{15\sqrt{15}}\frac{\sqrt{\alpha}^3}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1] \\ + \left[-\frac{\sqrt{\pi/6}}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}, \frac{1}{12n\hat{\theta}(1-\hat{\theta})}\right].\end{aligned} \tag{2.102}$$

The sum of the first three terms evaluates to $q(\tilde{\gamma})$. The remaining terms are now independent of γ and are of order $1/\sqrt{n}$. They can be merged by means of common bounds using $2n\hat{\theta}(1-\hat{\theta}) \geq \sqrt{n\hat{\theta}(1-\hat{\theta})}$, since $n\hat{\theta}(1-\hat{\theta}) \geq 1/2$ for our standing assumptions that $n \geq 1$ and $\hat{\theta}n$ is an integer different from 0 and n. Consequently, $12n\hat{\theta}(1-\hat{\theta}) \geq 6\sqrt{n\hat{\theta}(1-\hat{\theta})} \geq \sqrt{6/\pi}\sqrt{n\hat{\theta}(1-\hat{\theta})}$. The interval bounds then combine conservatively to

$$\frac{\sqrt{\pi/6} + 8\sqrt{\alpha}/\sqrt{15} + 64\sqrt{\alpha}^3/(15\sqrt{15})}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}. \tag{2.103}$$

We can now write

$$\alpha \in q(\tilde{\gamma}) + \frac{\sqrt{\pi/6} + 8\sqrt{\alpha}/\sqrt{15} + 64\sqrt{\alpha}^3/(15\sqrt{15})}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1], \tag{2.104}$$

which holds iff

$$q(\tilde{\gamma}) \in \alpha\left(1 + \frac{64\sqrt{\alpha}/(15\sqrt{15})}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1]\right) + \frac{\sqrt{\pi/6} + 8\sqrt{\alpha}/\sqrt{15}}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1]. \tag{2.105}$$

By monotonicity of q and extending q^{-1} to negative arguments as mentioned in the statement of Thm. 14 if necessary, the constraint is equivalent to

$$\tilde{\gamma} \in q^{-1}\left(\alpha\left(1 + \frac{64\sqrt{\alpha}/(15\sqrt{15})}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1]\right) + \frac{\sqrt{\pi/6} + 8\sqrt{\alpha}/\sqrt{15}}{\sqrt{n\hat{\theta}(1-\hat{\theta})}}[-1,1]\right). \tag{2.106}$$

For $\alpha > \log(2)$, we know that $\tilde{\gamma} > 0$, so we can add $\max(0, \ldots)$ as in the theorem statement.

To determine the interval equation for γ, we have $\gamma = \tilde{\gamma}\sqrt{\varphi(1-\varphi)/(\hat{\theta}(1-\hat{\theta}))}$. We use the first-order remainder to bound the factor on the right-hand side. For this consider the numerator,

and write $g(x) = \sqrt{(\hat{\theta} - x)(1 - \hat{\theta} + x)}$ with $0 \leq x \leq \hat{\theta} - \varphi$. We have

$$g^{(1)}(x) = \frac{2(\hat{\theta} - x) - 1}{2\sqrt{(\hat{\theta} - x)(1 - \hat{\theta} + x)}}, \tag{2.107}$$

$$|g^{(1)}(x)| \leq \frac{1}{2\sqrt{a_1 \hat{\theta}(1 - \hat{\theta})}}$$

$$= \frac{1}{\sqrt{3\hat{\theta}(1 - \hat{\theta})}}, \tag{2.108}$$

$$g(x) \in \sqrt{\hat{\theta}(1 - \hat{\theta})} + \frac{x}{\sqrt{3\hat{\theta}(1 - \hat{\theta})}}[-1, 1]. \tag{2.109}$$

With $x = \gamma\sqrt{\hat{\theta}(1 - \hat{\theta})/n}$ and the bound of $\gamma \leq 2\sqrt{3/5}\sqrt{\alpha}$, we get

$$\gamma \in \tilde{\gamma}\left(1 + \frac{2\sqrt{\alpha}/\sqrt{5}}{\sqrt{n\hat{\theta}(1 - \hat{\theta})}}[-1, 1]\right). \tag{2.110}$$

The theorem follows by composing this constraint with Eq. 2.106. □

Chapter 3

Deterministic drift instability and stochastic thermal perturbations of magnetic dissipative droplet solitons

This chapter written in collaboration with Mark Hoefer, [1] and Ezio Iacocca. [2] [3]

3.1 Introduction

Localized magnetic textures have recently attracted significant research interest due to their potential application in logic, storage, and communication technologies. From the perspective of logic and storage, static skyrmions [112] are very interesting textures due to their topological protection against perturbations, small sizes, and controllable motion [124, 128]. On the other hand, communication applications could benefit from dynamical textures, notably topological, dynamical kyrmions [165] and nontopological, magnetic dissipative droplets [66, 68, 100, 23, 73].

Magnetic dissipative droplets ("droplets" hereafter) have been widely observed in experiments both at cryogenic [98] and room temperatures [105, 9, 92, 36]. Droplets exist in magnetic thin films composed of materials with perpendicular magnetic anisotropy (PMA) [28], **i.e.**, in which the easy axis lies normal to the plane, so that it balances the exchange energy in favor of a localized structure [81, 66], Fig. 3.1(a). Furthermore, magnetic damping must also be balanced in order to sustain the droplet in time due to its lack of topology. To date, this has been achieved by using spin transfer torque (STT) [19, 135] in devices known as nanocontact spin torque oscillators, NC-

[1] Department of Applied Mathematics, University of Colorado, Boulder, CO, 80309
[2] Department of Applied Mathematics, University of Colorado, Boulder, CO, 80309
[3] Department of Applied Physics, Division for condensed matter theory, Chalmers University of Technology, 412 96, Gothenburg, Sweden

STOs [46]. NC-STOs are composed of a pseudo spin valve where two magnetic layers are decoupled by a nonmagnetic spacer, as shown in the schematic of Fig. 3.1(b). The topmost magnetic layer $\mathbf{m}$ is where the droplet nucleates and it is usually referred to as the free layer. The bottom magnetic layer $\mathbf{m}_\mathrm{p}$ serves as a spin-polarizer and it is known as the polarizer or fixed layer. In order to achieve sufficient current density to oppose magnetic damping, a nanocontact (NC) of radius R_* is placed on top of the free layer, confining the current to flow in an approximately cylindrical path [117] and therefore defining a region of effectively zero damping in the free layer. An external, perpendicular applied field H_0 is generally used in NC-STOs both to tilt the polarizer (useful for increasing STT and magnetoresistance), to provide an external source for the Larmor frequency, and to stabilize the droplet [23].

Since the first experimental observation of droplets [105], recent results have investigated theoretical predictions [66, 68, 67, 23, 24], shown the existence of hysteresis both at room and cryogenic temperatures [98, 92], identified a well-defined nucleation boundary [36], and even imaged the droplet via x-ray magnetic circular dichroism (XMCD) [9]. The same studies have demonstrated the existence of characteristics consistent with random droplet dynamics, notably low-frequency spectral features. These have been associated with the droplet exiting the NC region and succumbing to damping, a drift instability, originating from the spatial energy landscape created by the current-induced Oersted field [66, 117, 45, 99] and externally applied fields [23, 105, 24] or fluctuations in the material anisotropy spatial distribution [92]. However, the relationships between drift instabilities and physical sources of randomness have not been established. To provide an analytical understanding of drift instabilities, we study the effect of thermal noise on droplet dynamics.

In this chapter, we develop the stochastic evolution of droplet dynamics based on soliton perturbation theory [24] and obtain statistical observables such as the droplet center variance and the generation linewidth [134]. These results are analytically obtained by linearizing the equations of motion. From the linearization, we uncover a deterministic regime of drift instability, missed by previous analytical works [23, 24], where high bias currents induce growth of the droplet velocity on a long timescale. Randomness can also cause an otherwise deterministically stable

droplet to be expelled from the NC region when thermal fluctuations are taken into account. We determine that such events are extremely rare relative to the precessional timescale (10-100 picoseconds) but become quite relevant for the typical time scales of experiments (seconds or more). Observation, let alone quantification, of both the deterministic drift instability and the stochastic rare events is practically unfeasible utilizing standard deterministic [66] or stochastic [92] micromagnetic simulations alone. For a stable droplet, the generation linewidth is found to be dominated by the phase noise induced by a Wiener process or random walk, linearly proportional to temperature and inversely proportional to the NC radius. The droplet's center can be described by an Ornstein-Uhlenbeck (O-U) process with STT acting as an attractive mechanism that draws the droplet to the center of the NC. The determination of both stochastic processes requires subtle higher order effects from soliton perturbation theory [23, 24]. Full-scale micromagnetic simulations qualitatively agree with the analytical results, even when the current-induced Oersted field is taken into account.

The chapter is organized as follows. Section II describes the formalism used to obtain the stochastic equations for droplet dynamics. Section III explores the deterministic linearization where we obtain the fundamental droplet dynamical state and linear stability conditions. Stochastic terms are incorporated into the analysis in Sec. IV, leading to analytical solutions for the droplet center variance and generation linewidth at low temperatures. Numerical simulations of the nonlinear stochastic system are presented in Sec. V, demonstrating excellent agreement with the linearized analytical results. Full-scale micromagnetic simulations are used to explore regimes of small NC radii, nonlocal dipole fields, and Oersted field, beyond the scope of the asymptotic theory, nevertheless demonstrating qualitative agreement. Finally, we provide a discussion and concluding remarks in Sec. VI.

3.2 Droplet perturbation theory

The analytical study of droplet dynamics can be approached using perturbation theory with the magnetic damping and STT coefficients assumed small. This assumption alone yields droplet

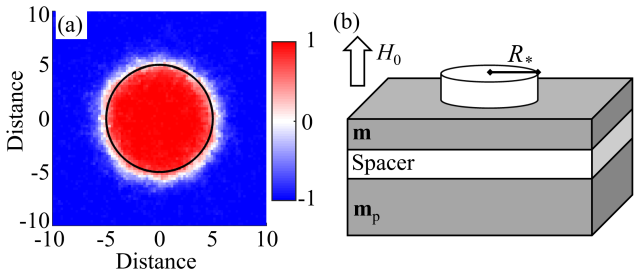

Figure 3.1: (a) Typical dissipative droplet obtained from micromagnetic simulations at a finite temperature. The $\hat{z}$ component of the magnetization is quantified by the color scale. (b) Schematic of a NC-STO based on a pseudo spin valve trilayer. The free, $\mathbf{m}$, and polarizer, $\mathbf{m}_\mathrm{p}$, magnetic layers are decoupled by a nonmagnetic spacer. A NC of radius R_* is placed on top of the free layer to achieve high current densities. An external field H_0 is applied perpendicularly to the plane.

nucleation conditions and the resultant droplet's frequency tunability via current and field [66, 36]. A semi-analytical generalization can be used to describe coarse droplet motion and control [67]. The additional assumption of a sufficiently large NC diameter implies a slowly precessing, circular domain wall description for the droplet [23], which enables a detailed analytical description of droplet dynamics in the presence of physical perturbation [23, 24]. This latter regime is the one considered here.

The equation of motion for the free layer magnetization $\mathbf{m}$ is the Landau-Lifshitz equation for a thin, two-dimensional magnetic film

$$\frac{\partial \mathbf{m}}{\partial t} = -\mathbf{m} \times \mathbf{h}_\mathrm{eff} + \mathbf{p}, \quad \mathbf{m} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{S}^2, \tag{3.1}$$

expressed here in nondimensional form. The effective field,

$$\mathbf{h}_\mathrm{eff} = h_0 \mathbf{z} + \nabla^2 \mathbf{m} + m_z \mathbf{z}, \tag{3.2}$$

includes contributions from a perpendicular external field h_0, the exchange field $\nabla^2 \mathbf{m}$, and a perpendicular magnetic anisotropy (PMA) field sufficient to overcome the thin-film limit of the demagnetizing field. Hence the $m_z \mathbf{z}$ term in the effective field has a positive coefficient, here scaled to unity. This form of the LL equation, with $|\mathbf{m}| = 1$, uses the time scale $\tau = (|\gamma| \mu_0 \eta)^{-1}$, where γ is the gyromagnetic ratio, μ_0 is the vacuum permeability, $\eta = M_s(Q-1)$ is the field scaling, M_s is the free layer's saturation magnetization, $Q = H_k/M_s$ is the nondimensionalized form of the PMA field

H_k, and the length scale $L = \lambda_{\text{ex}}/\sqrt{Q-1}$ where λ_{ex} is the exchange length. The NC radius R_* is nondimensionalized to $\rho_* = R_*/L$. We consider a small perturbation $|\mathbf{p}| \ll 1$ satisfying $\mathbf{p} \cdot \mathbf{m} = 0$ in order to preserve constant magnetization magnitude. The perturbation term considered here includes damping, STT as imposed by a NC-STO, and a thermal random field [27]

$$\mathbf{p} = \underbrace{-\alpha \mathbf{m} \times (\mathbf{m} \times \mathbf{h}_{\text{eff}})}_{\text{damping}} + \underbrace{\sigma \mathcal{H}(\mathbf{x}) \mathbf{m} \times \mathbf{m} \times \mathbf{m}_{\text{p}}}_{\text{NC-STO}} - \underbrace{\mathbf{m} \times \mathbf{h}}_{\text{thermal}}, \tag{3.3}$$

where $0 < \alpha \ll 1$ is the damping parameter,

$$\mathcal{H}(\mathbf{x}) = \begin{cases} 1 & |\mathbf{x}| \leq \rho_* \\ 0 & \text{else} \end{cases}$$

is a shifted Heaviside function describing the current path below the NC, $\mathbf{m}_{\text{p}}$ is the normalized polarizer orientation, and $\sigma = I/I_0$ is the nondimensionalized form of the current I, scaled by

$$I_0 = \frac{4\mu_0 M_s^2 (Q-1) e \pi R_*^2 \delta}{\hbar \epsilon}. \tag{3.4}$$

Here, e is the charge of the electron, δ is the thickness of the free layer, ϵ is the spin torque efficiency, and $\hbar$ is Planck's constant. Example scalings for recent experiments are listed in Table 3.1 for reference.

The thermal field $\mathbf{h}(\mathbf{x}, t)$ induces random fluctuations in the magnetization of a small material volume V and is assumed to be delta-correlated in space and time **i.e.**, white noise [27]. The variance

Parameters	Refs. [98] and [9]	Ref. [105]
τ (time, ns)	0.13	0.083
L (length, nm)	13.2	9.25
η (field, kA/m)	198.9	318.1
I_0(current, mA)	152.75	139.7
T_0(temperature, kK)	156.0	337.2
σ/α (scaled current to damping)	1.96	6.44
h_0 (scaled applied field)	0.5	2.5
ρ_* (scaled nanocontact radius)	5.96	5.95

Table 3.1: Time, length, field, current, and temperature scalings and typical nondimensionalized experimental parameters for recent experiments.

of the nondimensional field is $\mathrm{Var}[\mathbf{h}(\mathbf{x},t)] = \beta^2$ with

$$\beta^2 = \frac{T}{T_0}, \quad T_0 = \frac{\mu_0 M_s^2 V}{2\alpha k_B}, \tag{3.5}$$

where k_B is the Boltzmann constant, $V = \lambda_{ex}^2 \delta$ is the characteristic micromagnetic volume, and T_0 is the nondimensional scaling of the absolute temperature T. Table 3.1 includes typical temperature scalings for recent experiments and our micromagnetic simulations. The perturbative theory utilized here is valid in the low temperature regime where $\beta \ll 1$. The variance in Eq. (3.5) can be dimensionalized by multiplying (3.5) by $\tau M_s^2 (Q-1)^2$.

The droplet is characterized by its center position $\boldsymbol{\xi}$, velocity $\mathbf{v}$, collective phase ϕ, and precessional frequency ω. In the regime $0 \leq v \ll \omega \ll 1$, where $v = |\mathbf{v}|$ is the droplet speed, the droplet takes on the approximate form of a slowly precessing circular domain wall with a spatial phase proportional to the droplet's speed[24]

$$\cos\Theta = \tanh\left(\rho - \frac{1}{\omega}\right), \tag{3.6a}$$

$$\Phi = h_0 t - \frac{\mathbf{v}\cdot\hat{\boldsymbol{\rho}}}{\omega^2} + \phi, \quad \phi = \omega t + \phi_0. \tag{3.6b}$$

Equation (3.6) describes the magnetization orientation of the droplet in spherical coordinates (Θ, Φ) with polar angle from vertical $0 \leq \Theta < \pi$ and azimuthal angle Φ. In Eq. (3.6), we employ droplet-centered polar coordinates in the plane, so that the radial unit vector $\hat{\boldsymbol{\rho}}$ points from the droplet center $\boldsymbol{\xi}$ to a point in space $\mathbf{x} = (\rho\cos\varphi, \rho\sin\varphi)$ and the angular unit vector $\hat{\boldsymbol{\varphi}}$ is orthogonal $\hat{\boldsymbol{\varphi}}\cdot\hat{\boldsymbol{\rho}} = 0$ and satisfies the right hand rule $\hat{\boldsymbol{\rho}}\times\hat{\boldsymbol{\varphi}} = \mathbf{z}$. The structure of the $v = 0$ approximate droplet in Eq. (3.6) has been known for some time in the absence of STT (see, e.g., Ref. [81]) and the singular $\omega \to 0$ behavior for $v \neq 0$ was identified in Ref. [75].

Following the procedure described in Ref. [24], the slow evolution of the perturbed droplet's parameters for large NC radii $\rho_* \gg 1$, weak damping/STT $\sigma = \mathcal{O}(\alpha) \ll 1$, and low temperature

$\beta \ll 1$ is governed by the set of coupled, stochastic differential equations

$$d\phi = \omega\, dt - \frac{\sigma}{4\pi}\int_{|\mathbf{x}|\le\rho_*} (\mathbf{v}\cdot\hat{\boldsymbol{\rho}})\,\mathrm{sech}^2\left(\rho - \frac{1}{\omega}\right)\, d\mathbf{x}dt + d\mathcal{W}_\phi, \tag{3.7a}$$

$$d\boldsymbol{\xi} = \mathbf{v}dt + \frac{\sigma\omega}{2\pi}\int_{|\mathbf{x}|\le\rho_*} \mathrm{sech}^2\left(\rho - \frac{1}{\omega}\right)\hat{\boldsymbol{\rho}}\, d\mathbf{x}dt + d\boldsymbol{\mathcal{W}}_{\boldsymbol{\xi}}, \tag{3.7b}$$

$$d\omega = \alpha\omega^2(\omega + h_0)dt - \frac{\sigma\omega^3}{4\pi}\int_{|\mathbf{x}|\le\rho_*} \mathrm{sech}^2\left(\rho - \frac{1}{\omega}\right)\, d\mathbf{x}dt + d\mathcal{W}_\omega, \tag{3.7c}$$

$$d\mathbf{v} = \alpha\omega\mathbf{v}(\omega + 2h_0)dt - \frac{\sigma\omega^2}{2\pi}\int_{|\mathbf{x}|\le\rho_*} \left(\frac{3}{2}\mathbf{v} - \frac{(\mathbf{v}\cdot\hat{\boldsymbol{\varphi}})}{\rho\omega}\hat{\boldsymbol{\varphi}}\right)\mathrm{sech}^2\left(\rho - \frac{1}{\omega}\right)\, d\mathbf{x}dt + d\boldsymbol{\mathcal{W}}_{\mathbf{v}}, \tag{3.7d}$$

which are to be interpreted in the Stratonovich sense [39]. The dynamical system describing modulations of a droplet's parameters subject to a general class of perturbations was derived in Ref. [24]. The particular form of the stochastic terms in (3.7) result from the thermal field perturbation $-\mathbf{m}\times\mathbf{h}$ in eq. (3.3). The damping (proportional to α) and spin torque (proportional to σ) terms in (3.7) were also specifically derived in Ref. [24].

There is a symmetry in these equations between the droplet's collective precessional dynamics and motion. The phase ϕ and position $\boldsymbol{\xi}$ dynamics have a leading order linear coupling to the frequency ω and velocity $\mathbf{v}$ equations, respectively. The second, additional terms in the phase and position equations proportional to current σ correspond to higher order corrections from soliton perturbation theory, which prove to be essential for describing perturbed droplet dynamics [23, 24], in particular, the finite temperature effects explored here.

The terms $\mathcal{W}_i$, with $i = \phi$, $\boldsymbol{\xi}$, ω, and $\mathbf{v}$, are scaled Weiner processes, with nontrivial covariance structure. Each noise term is a spatial integral of the thermal field perturbation against an appropriate kernel (see Ref. [24], Eqs. 4.1–4.4). If we arrange them into a vector $\boldsymbol{\mathcal{W}} = \left(\mathcal{W}_\phi\ \mathcal{W}_{\xi_x}\ \mathcal{W}_{\xi_y}\ \mathcal{W}_\omega\ \mathcal{W}_{v_x}\ \mathcal{W}_{v_y}\right)^T$, then the covariance between the processes is given by

$$\mathrm{E}[\boldsymbol{\mathcal{W}}\boldsymbol{\mathcal{W}}^T] = \frac{\beta^2 t}{2\pi}\begin{bmatrix} \sigma_\phi^2 & \mathbf{v}^T/2 & & \mathbf{0} & \\ \mathbf{v}/2 & \omega\mathbf{I} & & & \\ & & \omega^5/2 & \omega^4\mathbf{v}^T & \\ & \mathbf{0} & & & \\ & & \omega^4\mathbf{v} & \boldsymbol{\sigma}_{\mathbf{v}}^2 & \end{bmatrix}, \tag{3.8}$$

where we have, for the sake of compactness, denoted

$$\sigma_\phi^2 = \frac{v^2}{4\omega} + \frac{\omega}{2}, \tag{3.9a}$$

$$\boldsymbol{\sigma}_{\mathbf{v}}^2 = \begin{bmatrix} \omega^5 + \frac{\omega^3}{4}\left(9v_x^2 + v_y^2\right) & 2\omega^3 v_x v_y \\ 2\omega^3 v_x v_y & \omega^5 + \frac{\omega^3}{4}\left(v_x^2 + 9v_y^2\right) \end{bmatrix} \tag{3.9b}$$

and $\mathbf{I}$ is the 2×2 identity matrix and $\mathbf{0}$ is the 3×3 zero matrix.

3.3 Deterministic Linearization and Stability

We will first examine the dynamics of Eqs. (3.7) at zero temperature $\beta^2 = 0$. These deterministic dynamics have been studied in detail [23, 24]. When the current σ exceeds the minimal sustaining current $\sigma_{\min}$, the system undergoes a saddle-node bifurcation resulting in a stable fixed point denoted $(\boldsymbol{\xi}_*, \omega_*, \mathbf{v}_*)$ that encapsulates the balance between damping and STT to sustain the droplet. The stable fixed point is stationary at the center of the NC, $\boldsymbol{\xi}_* = \mathbf{v}_* = 0$, with precessional frequency ω_* determined as a root of the transcendental equation

$$\frac{\sigma}{\alpha} = \frac{2(h_0 + \omega_*)}{1 + \omega_* \left[\log\left(\frac{1}{2}\mathrm{sech}\left(\rho_* - \frac{1}{\omega_*}\right)\right) + \rho_* \tanh\left(\rho_* - \frac{1}{\omega_*}\right)\right]}. \tag{3.10}$$

We observe that the phase ϕ in Eqs. (3.7) decouples from the system, so its dynamics can be determined from the remaining three parameters. If we linearize Eqs. (3.7) around this fixed point, we arrive at the system

$$\dot{\phi} = \omega, \tag{3.11a}$$

$$\dot{\boldsymbol{\xi}} = \mathbf{v} + \lambda_\xi \boldsymbol{\xi}, \tag{3.11b}$$

$$\dot{\omega} = \lambda_\omega(\omega - \omega_*), \tag{3.11c}$$

$$\dot{\mathbf{v}} = \lambda_v \mathbf{v}, \tag{3.11d}$$

where

$$\lambda_\xi = -\frac{1}{2}\sigma\rho_*\omega_*\mathrm{sech}^2\left(\rho_* - \frac{1}{\omega_*}\right), \tag{3.12a}$$

$$\lambda_\omega = -h_0\alpha\omega_* + \lambda_\xi + \frac{1}{2}\sigma\omega_*\left(\tanh\left(\rho_* - \frac{1}{\omega_*}\right) + 1\right), \tag{3.12b}$$

$$\lambda_v = -2\alpha\omega_*^2 + \lambda_\omega - \lambda_\xi. \tag{3.12c}$$

It is necessary to carefully choose parameters so that this fixed point is stable, **i.e.**, so that all eigenvalues in Eq. (3.12) are negative. The condition $\sigma > \alpha h_0$ is sufficient for $\lambda_\xi, \lambda_\omega < 0$, but in order to ensure that $\lambda_v < 0$, we require additionally that

$$\alpha\omega_*(2\omega_* + h_0) > \frac{1}{2}\sigma\omega_*\left(\tanh\left(\rho_* - \frac{1}{\omega_*}\right) + 1\right). \tag{3.13}$$

Note that the inequality requirement for stability in (3.13) was not identified previously [24], and is essential to understanding the dynamics of the droplet. It is possible to visualize the region of linear stability in the (h_0,σ/α) plane as in Fig. 3.2. The left (red) area corresponds to the condition $\sigma < \sigma_{\min}$, where the droplet cannot exist. This approximately linear relation for the existence boundary has been corroborated by experiment [98]. The inequality requirement Eq. (3.13) adds an unstable, right region (blue area) where the velocity of the droplet increases until it drifts away from the NC area and damping destroys it. The remaining white area represents the parameter space where the droplet exists and is stable. We observe that such a region shifts to lower applied fields and increased current for smaller NC radii (dashed lines).

It is helpful to express these eigenvalues in a more tractable form, so that we can observe how they approximately scale with experimental parameters. For this, we define a parameter-dependent constant that we will denote by ζ

$$\zeta = \frac{2\alpha h_0}{\sigma} - 1. \tag{3.14}$$

Previous work [24] assumed that the current was near the critical value $\sigma \approx 2\alpha h_0$, so that $\zeta = \mathcal{O}(\rho_*^{-1})$. This work relaxes that assumption and allows for any current that is sufficiently above the minimum sustaining current so that Eq. (3.10) can be approximately inverted to obtain the

frequency tunability

$$\omega_* = \rho_*^{-1} + \operatorname{arctanh}(\zeta)\rho_*^{-2} + \mathcal{O}(\rho_*^{-3}). \tag{3.15}$$

Then the leading-order approximations of each eigenvalue are

$$\lambda_\xi = -\frac{\sigma}{2}(1-\zeta^2) + \mathcal{O}(\rho_*^{-1}), \tag{3.16a}$$

$$\lambda_\omega = -\frac{\sigma}{2}(1-\zeta^2) + \mathcal{O}(\rho_*^{-1}), \tag{3.16b}$$

$$\lambda_v = \mathcal{O}(\sigma\rho_*^{-2}). \tag{3.16c}$$

The approximate expression for λ_v is prohibitively complex, so we omit it here in favor of the exact expression in eq. (3.12).

Inequality (3.13) is a fundamental result identifying a deterministic mechanism that can drive droplet drift instability. Any nonzero $\mathbf{v}$ (recall that $\mathbf{v} \neq 0$ corresponds to a spatial phase gradient across the droplet in Eq. (5.4)) will slowly increase when Eq. (3.13) does not hold. Large applied current destabilizes the droplet.

Previous work [23] that analyzed the dynamics of this system with $\mathbf{v} \equiv 0$, found that the dissipative droplet is linearly stable for physically relevant parameters. The sole effect of damping on a stationary droplet's frequency $d\omega/dt = \alpha\omega^3$ has been known for some time [14]. While this result demonstrates that the droplet is unstable in the presence of damping alone, it does not describe the instability investigated here with STT included. Similarly, Ref. [67] observed instability in the case $v \neq 0$ without taking into account STT. Our contribution here is to extend the dynamics to incorporate STT, thereby uncovering current dependent regimes of deterministic instability. In Ref. [24], the dynamics are analyzed, but the linear instability condition (3.13) was not recognized. The recognition of this linear instability, occurring at physically relevant parameters, is essential to the understanding of droplet dynamics. When $\mathbf{v} \neq 0$, the dynamics are much more sensitive to the choice of parameters, as is seen both above in the linear case, and below in the full nonlinear case. A key observation is that while $\lambda_\xi, \lambda_\omega$ are small $\mathcal{O}(\sigma)$, λ_v is much smaller $\mathcal{O}(\sigma\rho_*^{-2})$. These eigenvalues dictate the relaxation rate of the system towards the fixed point.

When compared to the $\mathbf{v} = 0$ dynamics, the relaxation rate of the droplet center $\boldsymbol{\xi}$ decreases by a factor proportional to $\rho_*^{-2} \ll 1$ when the $\mathbf{v}$ dynamics are included. Furthermore, we see that λ_v can change sign, while λ_ξ and λ_ω are negative for $\sigma > \alpha h_0$. This suggests that there is a shallow basin of attraction for the fixed point, allowing for the possibility of linear drift instabilities mediated by thermal noise. Indeed, all experiments [105, 98, 9] have been performed outside the region of linear stability, suggesting droplet drift instability and the concomitant observation of low-frequency spectral features. We note that the theory presented here is nominally applicable to the case $\rho_* \gg 1$, whereas the experiments in Refs. [105, 98, 9] with $\rho_* \in (5, 8)$ are at the borderline of applicability.

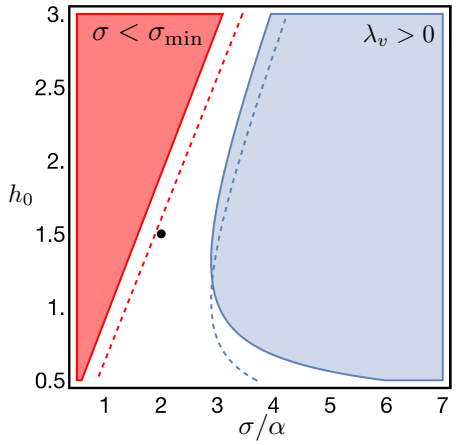

Figure 3.2: Droplet existence and linearly stable parameter space for a droplet nucleated in a NC of normalized radius $\rho_* = 15$. The droplet cannot exist in the left region (filled red) where $\sigma < \sigma_{\min}$, whereas the droplet is linearly unstable in the right region (filled blue) where $\lambda_v > 0$. Therefore, the droplet is stable in the remaining white region. The numerical simulations in Sec. 3.5 and Figs. 3.3, 3.4, 3.5(a) with $h_0 = 1.5$, $\alpha = 0.03$, and $\sigma = 2\alpha$ (black circle) exhibit linear stability. The dashed lines are boundaries for droplet existence and linear stability with reduced NC radius $\rho_* = 5$.

3.4 Stochastic linearization

One method of approximating the dynamics of the stochastic system Eqs. (3.7) is to employ the previously calculated linearization of the deterministic system and approximate the noise, now denoted $\boldsymbol{\mathcal{W}}^*$, by evaluating the covariance matrix, Eq. (3.8), at the fixed point. This low

temperature theory yields the linear stochastic system

$$d\phi = \omega dt + d\mathcal{W}^*_\phi, \tag{3.17a}$$

$$d\boldsymbol{\xi} = \mathbf{v}dt + \lambda_\xi \boldsymbol{\xi} dt + d\boldsymbol{\mathcal{W}}^*_{\boldsymbol{\xi}}, \tag{3.17b}$$

$$d\omega = \lambda_\omega(\omega - \omega_*)dt + d\mathcal{W}^*_\omega, \tag{3.17c}$$

$$d\mathbf{v} = \lambda_v \mathbf{v} dt + d\boldsymbol{\mathcal{W}}^*_{\mathbf{v}}. \tag{3.17d}$$

When evaluated at the fixed point, the covariance matrix becomes diagonal

$$\begin{aligned} \mathrm{E}[\boldsymbol{\mathcal{W}}^*\boldsymbol{\mathcal{W}}^{*\boldsymbol{T}}] &= \beta^2 t \cdot \mathrm{Diag}\left(\frac{\omega_*}{4\pi}, \frac{\omega_*}{2\pi}, \frac{\omega_*}{2\pi}, \frac{\omega_*^5}{4\pi}, \frac{\omega_*^5}{2\pi}, \frac{\omega_*^5}{2\pi}\right), \\ &\equiv t \cdot \mathrm{Diag}\left(\beta_\phi^2, \beta_\xi^2, \beta_\xi^2, \beta_\omega^2, \beta_v^2, \beta_v^2\right), \end{aligned} \tag{3.18}$$

where we denote the variance of each parameter by β_i^2 for $i = \phi$, $\boldsymbol{\xi}$, ω, and $\mathbf{v}$. The linear system Eqs. (3.17) can be solved explicitly. For $(\boldsymbol{\xi}, \omega, \mathbf{v})$, we obtain a set of coupled O-U processes that describe the stochastic properties of the linear system.

Of particular interest is the behavior of the decoupled oscillator phase, $\phi(t)$, as it allows us to relate our analytical description with the generation linewidth, Δf, which can be measured from the electrical characterization of NC-STOs [46]. Solving the system in Eqs. (3.17), we find that the frequency is an O-U process with mean ω_* and variance

$$\begin{aligned} \mathrm{Var}[\omega(t)] &= -\frac{\beta_\omega^2}{2\lambda_\omega}\left(1 - \exp 2\lambda_\omega t\right) \\ &\to -\frac{\beta_\omega^2}{2\lambda_\omega}, \quad \text{as } t \to \infty. \end{aligned} \tag{3.19}$$

We can then write down the solution for the phase $\phi(t)$ as the sum of an integrated O-U process and a Wiener process (random walk)

$$\phi(t) = \int_0^t \omega(s)ds + \mathcal{W}^*_\phi, \tag{3.20}$$

from which we find that the variance of ϕ quickly approaches linear growth

$$\mathrm{Var}[\phi(t)] \to \left(\frac{\beta_\omega^2}{\lambda_\omega^2} + \beta_\phi^2\right) t \quad \text{as } t \to \infty, \tag{3.21}$$

so that the spectral lineshape is Lorentzian and the generation linewidth is given by

$$\Delta f = \left(\frac{\beta_\omega^2}{\lambda_\omega^2} + \beta_\phi^2 \right) = \beta^2 \left(\frac{\omega_*^5}{4\pi\lambda_\omega^2} + \frac{\omega_*}{2\pi} \right). \tag{3.22}$$

By virtue of Eq. (3.5), the generation linewidth is linearly proportional to temperature. Because $\omega_* \ll 1$, the generation linewidth is dominated by the Wiener process, phase noise contribution [second term in Eq. (3.22)] resulting from the higher order contribution to the phase dynamics of Eq. (5.2). This expression for generation linewidth is also consistent with the notion of a reduced impact of thermal fluctuations on a larger magnetic mode volume. Indeed, recalling Eq. (5.7), it is clear that larger NC radii minimize the generation linewidth.

We are also interested in the dynamics of the center $\boldsymbol{\xi}$ as it describes the droplet's random motion with respect to the NC region. The velocity and position form a coupled pair of O-U processes, which we can solve using standard methods. We then find the variance of the droplet center

$$\begin{aligned} s_{\boldsymbol{\xi}}^2(t) = &-\frac{\beta_\xi^2}{2\lambda_\xi}\left(1 - e^{2\lambda_\xi t}\right) - \frac{\beta_v^2}{2\left(\lambda_\xi - \lambda_v\right)^2} \\ &\times \left(\frac{1 - e^{2\lambda_\xi t}}{\lambda_\xi} + \frac{4\left(1 - e^{\left(\lambda_\xi + \lambda_v\right)t}\right)}{\lambda_\xi + \lambda_v} + \frac{1 - e^{2\lambda_v t}}{\lambda_v} \right) \\ &\to -\frac{1}{2}\frac{\beta_v^2}{\lambda_\xi^2 \lambda_v} - \frac{\beta_\xi^2}{\lambda_\xi} \quad \text{as } t \to \infty. \end{aligned} \tag{3.23}$$

We might expect that the position noise term in Eq. (3.23) would dominate, analogous to the phase noise in Eq. (3.22). However, the balance of the two terms in Eq. (3.23) is highly sensitive to experimental parameters. In fact, for the parameters used in this study, the velocity noise term is the dominant contribution.

Equations (3.22) and (3.23) are central results of this chapter. The former relates our stochastic theory to an experimental observable, namely, the Lorentzian generation linewidth. The latter quantifies the amount of droplet drift with respect to the NC center and thus provides a means to quantify the drift instability from random fluctuations in the magnetic system.

3.5 Numerical Simulations

To examine the behavior of the full nonlinear system, Eqs. (3.7), we numerically simulate an ensemble of sample paths. Details of our numerical implementation can be found in the Appendix. We choose the parameters $\rho_* = 15$, $h_0 = 1.5$, $\alpha = 0.03$, and $\sigma = 2\alpha$ in order to ensure that we are within the asymptotic validity of our analysis and the region of linear stability, depicted by the black dot in Fig. 3.2. Typical sample paths of the droplet's phase and center position generated by this method are shown in Fig. 3.3.

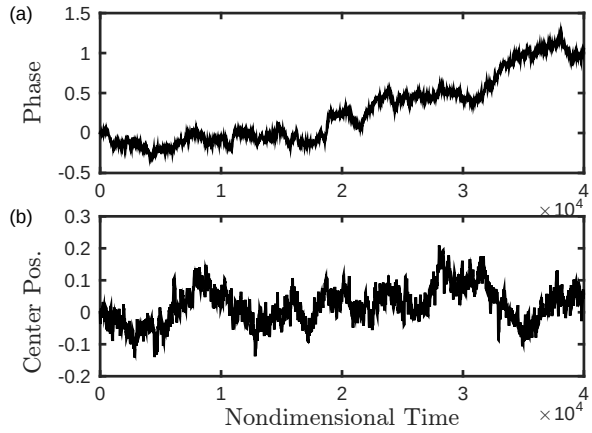

Figure 3.3: Numerically computed nonlinear sample path from Eqs. (3.7) for (a) the droplet phase ϕ and (b) the x-component of the center position ξ_x. The phase is measured in radians, and position and time are nondimensional as per Table 3.1.

We first examine the statistics of the droplet center. Figure 3.4 shows the standard deviation of the droplet center for an ensemble of numerical simulations of the linear (blue) and nonlinear (red) systems. The analytical prediction of Eq. (3.23) (black) agrees well with the linear simulation. For the chosen set of parameters, nonlinearity is not observed to significantly enhance the droplet drift and, in fact, the standard deviation of the droplet center from the NC center is never more than 1% of the NC radius. For slightly modified parameters, we observe qualitatively different behavior when nonlinearity is introduced. For example, reducing the NC radius to $\rho_* = 10$, we approach the stability boundary of Fig. 3.2, although linear stability in Eq. (3.13) is still satisfied.

However, the numerical simulation of the nonlinear system leads to approximately 15% of the simulated paths leaving the NC before $t = 4 \cdot 10^4$. This indicates that the basin of attraction of the system is relatively small. We can also infer from the simulations at larger NC radii that the size of the basin of attraction decreases with NC radius. This suggests that the small NC devices used in experiments at room temperature sustain droplets that exhibit deterministic or thermally induced drift instabilities during measurements. In fact, typical spectral measurements [105, 92] acquire data in time spans on the order of seconds, which translate to $\approx 1 \cdot 10^{10}$ in our normalized units. The characterization of the multi-dimensional boundary in phase space of the basin of attraction and ejection statistics are, however, outside the scope of this chapter. Note that in the ensemble used to generate Fig. 3.4, no sample paths ejected from the NC.

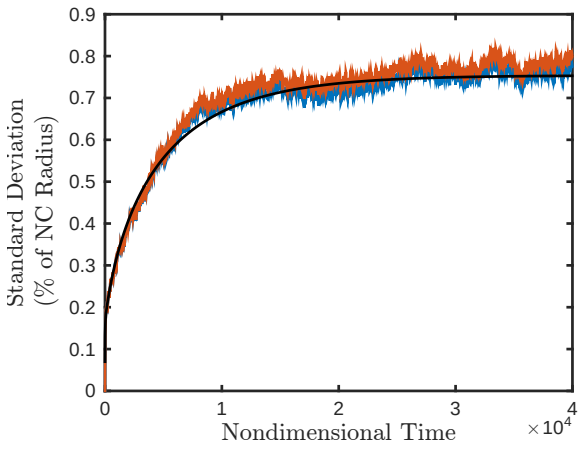

Figure 3.4: Standard deviation $s_{\boldsymbol{\xi}}$ of the droplet center from linear theory (solid black line), linear simulation (solid blue), and nonlinear simulation (solid red).

In the regime where the droplet does not drift away from the NC over the timescale simulated, it is possible to compare the linear generation linewidth to numerical simulations of Eqs. (3.7). From a sample path of the stochastic phase $\phi(t)$, we calculate the linewidth via the power spectral density of ϕ, as discussed in Ref. [134]. It is worth noting that the linewidth calculated via this method is strictly valid for white noise [47] and can vary between sample paths due to the fluctuations between each path. For the linewidths reported here, we take the mean value of the calculated linewidths

from 500 different sample paths. Figure 3.5(a) shows the linewidth's dependence on temperature for the nonlinear system (red asterisks) and Eq. (3.22) (solid black line). Finite sampling and the asymmetric, heavy-tailed distribution of linewidths across sample paths causes the mean to converge slowly to the linear theory at low temperature. Although the median gives results more clearly convergent to the linear theory, the mean corresponds to experimentally observed linewidths, which are averaged over long timescales. Nonlinear simulations yield a mean linewidth of $1.77 \cdot 10^{-5}$ at temperature $\beta^2 = 2.8 \cdot 10^{-3}$, which, for comparative purposes, corresponds to 214 kHz at temperature $T = 314$ K under the temporal and temperature scalings of Ref. [105] with damping enhanced to $\alpha = 0.03$. Note, however, that the material parameters (ρ_*, σ, and h_0) for the nonlinear simulations do not correspond to those from Ref. [105].

The linear theory is a very good predictor of the nonlinear system's behavior at low temperatures and we numerically observe that the discrepancy between the linear and nonlinear linewidths decreases quadratically in T as $T \to 0$, as one would expect from this perturbative approach. However, as room temperature is approached, the nonlinear linewidth exceeds the linear linewidth by an order of magnitude. This originates from the increased impact of thermal fluctuations when the linearization is not strictly applicable. We stress that current experiments have not directly detected ejection events, and in the event of ejection, the bias current can re-nucleate a droplet and the resulting linewidth in aggregate will be considerably broader due to the ensuing transient dynamics. Our simulations end upon ejection, and do not allow for re-nucleation.

The above simulations are strictly valid for the regime $\rho_* \gg 1$ with negligible long-range dipole and Oersted fields. Experiments to date, however, have been performed when $\rho_* \in (5, 8)$. Moreover, it is important to characterize the impact of dipolar and Oersted fields on the droplet's collective motion and precession. To further explore droplet behavior, we perform full-scale micromagnetic simulations with nonlocal dipole fields using the GPU-based package Mumax3 [147]. We first compare micromagnetic results by choosing the same set of dimensionless parameters specified above and scalings consistent with Co/Ni multilayers [105] (See Table 3.1). The fixed layer is assumed to be perpendicularly polarized. The NC is placed at the geometrical center of an active

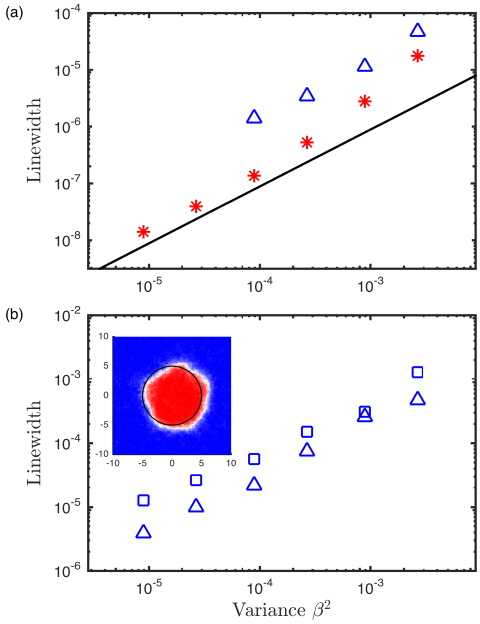

Figure 3.5: (a) Droplet linewidth as a function of temperature from linear theory (solid black line), nonlinear simulation (red asterisks), and micromagnetic simulations (blue triangles) when $\rho_* = 15$, $h_0 = 1.5$, $\alpha = 0.03$, $\sigma = 2\alpha$. (b) Droplet linewidth as a function of temperature from micromagnetic simulations where only the NC radius is reduced to $\rho_* = 5$ from (a) (blue triangles) and the effect of a current-induced Oersted field (blue squares). Linewidth is expressed in rad/τ as per Table 3.1. Inset shows droplet profile with Oersted field included. Error bars are $\mathcal{O}(10^{-9})$ and are not shown.

area of size $89.9 \times 89.9 \times 0.39$ discretized in cells with size $0.35 \times 0.35 \times 0.39$, below the exchange length. An ensemble of sample paths is not feasible to compute micromagnetically due to time constraints, so we determine the linewidth from a single path spanning $t = 1.8 \times 10^4$ and sampled at 0.015 intervals. First, we do not observe droplet motion, which is consistent with the results shown in Fig. 3.4 where the droplet center variance is expected to be below our cell resolution. The results for the temperature dependent linewidth are shown in Fig. 3.5(a) as blue triangles. We note that the micromagnetic simulations overestimate the linewidth obtained from nonlinear simulations but are on the same order of magnitude at room temperatures. At low temperatures, the micromagnetic simulations do not approach the linear theory as one would expect. This is a consequence of the limited simulation time and the spatial resolution of our micromagnetic scheme that precludes an accurate estimation of the phase noise statistics and thus its convergence to the linear linewidth.

Despite this limitation, micromagnetic simulations can be used to explore the dynamics of droplets sustained in devices with smaller NC radii, where micromagnetics have shown to be more accurate [105, 92] and where, conversely, the theory is not strictly applicable. We perform micromagnetic simulations with the same nondimensional parameters specified above but reduce the NC radius to $\rho_* = 5$, in the range of experiments performed to date, and increase the current to $\sigma = 0.1$. The resulting linewidths are shown in Fig. 3.5(b) by blue triangles. A qualitative agreement with theory is observed, namely, a linear dependence of the linewidth on temperature and a linewidth increase for smaller NC radii. Additionally, micromagnetic simulations allow us to include the current-generated Oersted field [117, 45, 99]. This further enhances the linewidth by a factor ~ 5 (blue squares) originating from the distortion of the droplet boundary, as observed from a snapshot of the $\hat{\mathbf{z}}$ magnetization component shown in the inset of Fig. 3.5(b). These results suggests that the unavoidable nonlocal and Oersted fields in a real device will enhance the generation linewidth compared to theory, but the temperature-dependent features remain mostly unchanged.

3.6 Discussion and Conclusion

We have developed a stochastic perturbation theory for magnetic dissipative droplets describing the random motion of the droplet's position, velocity, frequency, and phase. Higher-order perturbative effects in the phase and position are shown to be essential for understanding the dynamics of the droplet. Inclusion of velocity dynamics causes a qualitative shift in the behavior of the droplet position, and gives rise to a previously uncovered deterministic drift instability. Such an instability occurs at high driving currents, leading to an exponential increase in the droplet velocity. This effect also implies a small basin of attraction for the stable fixed point, providing a simple explanation for the origin of drift instabilities from randomness in the system, such as thermal fluctuations.

We find that in parameter regimes where the deterministic droplet is linearly stable, the stochastically induced drift instabilities are rare events compared with the typical precessional timescales. A notable implication is that the observation of drift instabilities due to thermal fluctuations using micromagnetic simulations is prohibitive. In contrast, our finite dimensional reduction of the governing partial differential equation makes such effects computationally feasible. The study of rare events is beyond the scope of this chapter, but motivates an application of large deviation theory, as previously studied, for example, in the context of fiber optic soliton communication systems [110]. Likewise, micromagnetic simulations tailored to study rare events [149] might be used to resolve the time and computational limitations. Even in the deterministic case, the predicted linear instability may be difficult to recover from micromagnetic simulations due to its slow rate of exponential growth. From an experimental point of view, typical measurement timescales suggest that drift instability and droplet renucleation can occur many times. For example, the long timescale required in the direct imaging of localized excitations, 500 ms, indicates that drift instabilities could occur $\sim 10^6$ times, leading to the small droplet amplitude and spatial smearing observed in the XMCD images of Ref. [9].

In contrast, previous works have interpreted the droplet drift mechanism through spatial

inhomogeneities in field [24] or anisotropy [92]. Here, we have identified two additional drift mechanisms, a deterministic linear instability inherent to the NC-STO system and rare drift events caused by thermal fluctuations.

Our model also allows us to obtain an analytical expression for the linearly stable droplet generation linewidth. At low temperature, we find that the phase noise is characterized by a Wiener process (random walk) and the droplet center is an O-U process, analogous to the stochastic phase and amplitude dynamics, respectively, of spatially uniform STOs [134]. For the linearized system, the resulting generation linewidth is linearly dependent on temperature, whereas the nonlinear system exhibits a linewidth enhancement when approaching room temperature, reflecting the coupling between the droplet's constituent variables. Full-scale micromagnetic simulation, including the fully nonlinear spatial variation of the system, qualitatively agree with the numerical results. However, we do not observe convergence toward the linear theory at low temperatures using a standard micromagnetic package [147]. This suggests the study of droplet generation linewidth as a test problem for stochastic micromagnetic codes [119].

The analytical and numerical linewidths obtained are two orders of magnitude below the typical linewidths observed in experiments. This disagreement may be caused by the small NC radii used experimentally, the existence of nonlocal dipolar and current-induced Oersted fields, and the aforementioned drift instabilities for data-acquisition timescales. In fact, micromagnetic simulations performed with a radius similar to those experimentally fabricated to date return linewidths in the same order of magnitude when both nonlocal and current-induced Oersted fields are included. The relevance of such fields in the generation linewidth motivates their inclusion in the analytical theory. For thin films, the effect of nonlocal dipole fields on deterministic droplet dynamics has been shown to be a frequency downshift when $\mathbf{v} = 0$ [23]. It remains to incorporate these effects into the stochastic theory when $\mathbf{v} \neq 0$. Because the Oersted field is not a singular perturbation [24], its inclusion in this collective theory would necessitate the incorporation of droplet coupling to spin waves. Such coupling is in principle possible, see, e.g., Ref. [77].

In [93], the authors perform an experimental examination of the effect of temperature on

droplet nucleation and stability. They find that droplet nucleation can occur at lower external field strength when the temperature is lower, as one would expect. In the current work, we consider the regimes of droplet stability with respect to zero temperature; this experimental understanding of the influence of temperature on the stable paramter regime is valuable, as it will help us further analyze the sources of observed droplet instability.

In conclusion, this work provides the means to seek optimized experimental parameters for a given application. To wit, we find that an environment with a large NC radius, low field, modest current, and large anisotropy are less susceptible to drift and thus lead to a much narrower generation linewidth. Our results motivate a more detailed experimental study on the current and temperature-dependent generation linewidth and ejection statistics of droplets.

3.7 Appendix

3.7.1 Numerical Methods

We simulate the nonlinear system Eq. (3.7) via the Euler-Maruyama method, with drift correction to account for the Stratonovich interpretation of the stochastic integrals [80]. Results of the higher-order Milstein scheme[80] yield negligibly different results. For example, the linewidth at room temperature differed by $< 1\%$ between the Euler and Milstein simulations. We use a timestep of $dt = 4$, and our total integration time is $t = 4 \cdot 10^4$. We integrate 500 sample paths, and then use the standard sample variance to produce Figure 3.4.

We must ensure that our nonlinear and linear systems coincide when $T \to 0$. To that end, we calculate the pathwise difference between the droplet center $\boldsymbol{\xi}_{\mathrm{L}}$ calculated by discretizing the linear system Eq. (3.17) and the droplet center $\boldsymbol{\xi}_{\mathrm{NL}}$ calculated via discretizing the nonlinear system Eq. (3.7). Note that both paths are calculated using the same stochastic terms, scaled appropriately. The results are shown in Figure 3.6. The standard deviation of the droplet center from the fixed point is $\mathcal{O}(\sqrt{T})$, and the separation between the nonlinear and linear paths is $\mathcal{O}(T)$, so we have

$$\|\boldsymbol{\xi}_{\mathrm{NL}}(t) - \boldsymbol{\xi}_{\mathrm{L}}(t)\| = \mathcal{O}(s_\xi^2). \tag{3.24}$$

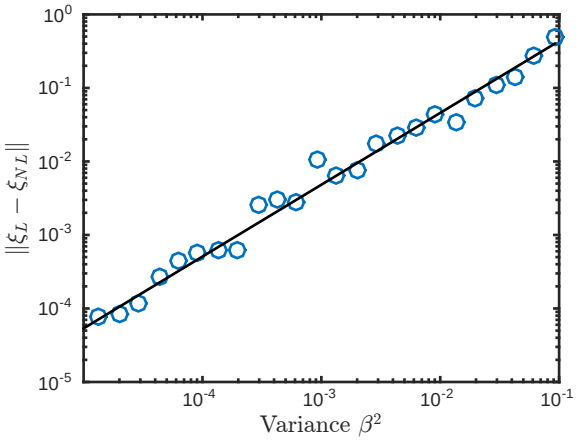

Figure 3.6: Convergence plot indicating that $|\xi_{\mathrm{NL}} - \xi_{\mathrm{L}}| = \mathcal{O}(\beta^2) = \mathcal{O}(T)$. Best-fit line has slope of 0.9793.

This linear convergence in T is a positive consistency check on the linearization Eq. (3.17) and stochastic timestepping of the nonlinear system Eq. (3.7).

Chapter 4

Detecting Topological Changes in Dynamic Community Networks

This chapter written in collaboration with François Meyer. [1] [2]

4.1 Introduction

The study of time-varying (dynamic) networks (or graphs) is of fundamental importance for computer network analytics and the detection of anomalies associated with cyber crime [70, 74, 78]. Dynamic graphs also provide models for social networks [5, 56], and are used to decode the functional connectivity in neuroscience [60, 72, 137] and biology [15]. The significance of this research topic has triggered much recent work [6, 88, 121]. Several methods have been proposed to detect the effect of significant structural changes (e.g., changes in topology, connectivity, or relative size of the communities in a community graph) in a time series of graphs. We focus on networks that change over time, allowing both edges and nodes to be added or removed. We refer to these as **dynamic networks**.

A fundamental goal of the study of dynamic graphs is the identification of universal patterns that uniquely couple the dynamical processes that drive the evolution of the connectivity with the specific topology of the network; in essence the discovery of universal spatio-temporal patterns [86, 76]. In this context, the goal of the present work is to detect anomalous time steps in the evolution of dynamic graphs. We propose a novel statistical method, which captures the coherence of the dynamics under baseline (normal) evolution of the graph, and can detect switching and

[1] Department of Electrical and Computer Engineering, University of Colorado, Boulder, CO, 80309
[2] Department of Applied Mathematics, University of Colorado, Boulder, CO, 80309

regime transitions triggered by anomalies. Specifically, we study a mathematical model of normal and abnormal growth of a community network. We are motivated to study this model because many real-world networks are of this type [113]. Members join and leave social networks; new neurons develop in the brain and form connections with existing structures.

In order to circumvent the problem of decomposing each graph into simpler structures (e.g., communities), we use a metric to quantify changes in the graph topology as a function of time. The detection of anomalies becomes one of testing the hypothesis that the graph is undergoing a significant structural change. Several notions of similarity have been proposed to quantify the structural similitude without resorting to the computation of a true distance (e.g., [16, 84] and references therein). Unlike a true metric, a similarity is typically not injective (two graphs can be perfectly similar without being the same), and rarely satisfies the triangular inequality. This approach relies on the construction of a feature vector that extracts a signature of the graph characteristics; the respective feature vectors of the two graphs are then compared using a distance, or a kernel. In the extensive review of Koutra et al. [84], the authors studied several graph similarities and distances. They concluded that existing similarities and distances either fail to conform to a small number of well-founded axioms, or suffer from a prohibitive computational cost. In response to these shortcomings, Koutra et al. proposed a novel notion of similarity [84].

Inspired by the work of [84], in [107] the authors propose a true metric that address some of the limitations of the DeltaCon similarity introduced in [84]. We emphasize that it is highly preferable to have a proper metric, rather than an informal distance, when comparing graphs; this allows one to employ proof techniques not available in the absence of the triangle inequality. Our distance, coined the **resistance-perturbation distance**, can quantify structural changes occurring on a graph at different scales: from the local scale formed by the neighbors of each vertex, to the largest scale that quantifies the connections between clusters, or communities. Furthermore, we proposed fast (linear in the number of edges) randomized algorithms that can quickly compute an approximation to the graph metric, for which error bounds are proven (in contrast to the `DeltaCon` algorithm given in [84], which has a linear time approximate algorithm but for which no error

bounds are given).

The main contribution of this work is a detailed analysis of a dynamic community graph model, which we call the dynamic stochastic blockmodel. This model is formed by adding new vertices, and randomly attaching them to the existing nodes. The goal of the work is to detect the time at which the graph dynamics switches from a normal evolution – where two balanced communities grow at the same rate – to an abnormal behavior – where the two communities are merging. Because the evolution of the graph is stochastic, one expects random fluctuations of the graph geometry. The challenge is to detect an anomalous event under normal random variation. We propose an hypothesis test to detect the abnormal growth of the balanced stochastic blockmodel. In addition to the theoretical analysis of the test statistic, we conduct several experiments on synthetic and real dynamic networks, and we demonstrate that our test can detect changes in graph topology.

The remainder of this chapter is organized as follows. In the next section we introduce the main mathematical concepts and corresponding nomenclature. In section 4.3 we recall the definition of the resistance perturbation distance. We provide a straightforward extension of the metric to graphs of different sizes and disconnected graphs. In section 4.4 we formally define the problem, we introduce the dynamic **balanced two-community stochastic blockmodel**, and we describe the main contributions and the line of attack. Finally, in Section 4.6 we use our methodology to detect significant changes in synthetic and real dynamic networks.

4.2 Preliminaries and Notation

We denote by $G = (V, E)$ an undirected, unweighted graph. We will often use u, v, or w to denote vertices in V. We use the standard asymptotic notation; see Appendix 4.8.1 for details. Given a family of probability spaces $\Omega = (\Omega_n, \mathrm{Prob}_n)$, and a sequence of events $E = (E_n)$, we write that Ω has the property **with high probability** ("w.h.p."), if $\lim_{n\to\infty} \mathrm{Prob}\,(E_n) = 1$. Finally, we use an abbreviated summation notation, where

$$\sum_{u\le n} \text{ is short for } \sum_{u=1}^{n} \text{ and } \sum_{u<v\le n} \text{ is short for } \sum_{u=1}^{n}\sum_{v=u+1}^{n}.$$

Table 5.5 in Appendix 4.8.2 provides a list of the main notations used in the chapter.

4.2.1 Effective Resistance

We briefly review the notion of effective resistance [79, 44, 55, 48] on a connected graph. The reader familiar with the concept can jump to the next section. There are many different ways to present the concept of effective resistance. We use the electrical analogy, which is very standard (e.g., [44]). Given a graph $G = (V, E)$, we transform G into a resistor network by replacing each edge e by a resistor with conductance w_e (i.e., with resistance $1/w_e$).

Definition 1 (Effective resistance [79])**.** *The effective resistance $\widehat{R}_{uv}$ between two vertices u and v in V is defined as the voltage applied between u and v that is required to maintain a unit current through the terminals formed by u and v.*

The relevance of the effective resistance in graph theory stems from the fact that it provides a distance on a graph [79] that quantifies the connectivity between any two vertices, not simply the length of the shortest path. Changes in effective resistance reveal structural changes occurring on a graph at different scales: from the local scale formed by the neighbors of each vertex, to the largest scale that quantifies the connections between clusters, or communities.

4.3 Resistance Metrics

4.3.1 The Resistance Perturbation Metric

The effective resistance can be used to track structural changes in a graph, and we use it to define a distance between two graphs on the same vertex set [107]. Formally, we define the Resistance Perturbation Distance as follows.

Definition 2 (Resistance Perturbation Distance)**.** *Let $G^{(1)} = (V, E^{(1)}, w^{(1)})$ and $G^{(2)} = (V, E^{(2)}, w^{(2)})$ be two connected, weighted, undirected graphs on the same vertex set, with respective effective resistance matrices, $\widehat{\boldsymbol{R}}^{(1)}$ and $\widehat{\boldsymbol{R}}^{(2)}$, respectively. The **RP-p distance** between $G^{(1)}$ and $G^{(2)}$ is*

defined as the element-wise p-norm of the difference between their effective resistance matrices. For $1 \leq p < \infty$,

$$\mathrm{RP}_p(G^{(1)}, G^{(2)}) = \left\| \widehat{\boldsymbol{R}}^{(1)} - \widehat{\boldsymbol{R}}^{(2)} \right\|_p = \left[\sum_{i,j \in V} \left| \widehat{R}_{ij}^{(1)} - \widehat{R}_{ij}^{(2)} \right|^p \right]^{1/p}. \tag{4.1}$$

In this chapter, we will restrict our attention to the RP_1 distance (we will omit the subscript $p = 1$), because it is directly analogous to the Kirchhoff index.

4.3.2 Extending the Metric to Disconnected Graphs

As defined, the resistance metric is not properly defined when the vertices are not within the same connected component. To remedy this, we use a standard approach. Letting $\widehat{R}_{uv}$ denote the effective resistance between two vertices u and v in a graph, then the conductivity $C_{uv} = \widehat{R}_{uv}^{-1}$ can be defined to be zero for vertices in disconnected components. Considering the conductivity as a similarity measure on vertices, a distance is given by the quantity $(1 + C_{uv})^{-1}$. Note that $(1 + C_{uv})^{-1} = \widehat{R}_{uv}/(\widehat{R}_{uv} + 1)$, and so we can define this new quantity relative to the effective resistance without any reference to the conductance. We refer to the resulting quantity as the **renormalized effective resistance**.

Definition 3 (Renormalized Effective Resistance)**.** *Let* $G = (V, E)$ *be a graph (possibly disconnected).*

We define the renormalized effective resistance between any two vertices u *and* v *to be*

$$R_{uv} = \begin{cases} \widehat{R}_{uv}/(\widehat{R}_{uv} + \beta) & \text{if } u \text{ and } v \text{ are connected,} \\ 1 & \text{otherwise,} \end{cases} \tag{4.2}$$

where $\widehat{R}_{uv}$ *is the effective resistance between* u *and* v*, and* $\beta > 0$ *is an arbitrary constant.*

We now proceed to extend the notion of resistance perturbation distance.

Definition 4 (Renormalized Resistance Distance)**.** *Let* $G_1 = (V_1, E_1)$ *and* $G_2 = (V_2, E_2)$ *be two graphs (with possibly different vertex sets). We consider* $V = V_1 \cup V_2$*, and relabel the union of*

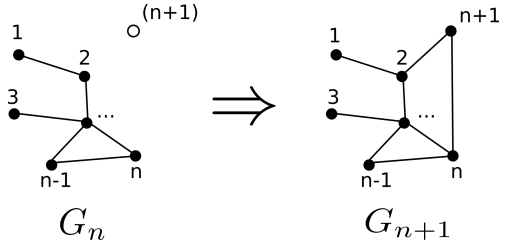

Figure 4.1: In order to compare G_n and G_{n+1}, we include node $n+1$ into G_n (see left), and evaluate the renormalized effective resistance on the augmented graph, with vertex set $\{1,\ldots,n\}\cup\{n+1\}$. Using definition 4, the two graphs G_n and G_{n+1} can then be compared.

vertices using $[n]$*, where* $n = |V|$*. Let* $R^{(1)}$ *and* $R^{(2)}$ *denote the renormalized effective resistances in* $\widetilde{G_1} = (V, E_1)$ *and* $\widetilde{G_2} = (V, E_2)$*, respectively.*

We define the renormalized resistance distance to be

$$\mathrm{RD}_\beta(G_1, G_2) = \sum_{u<v\le n} \left| R^{(1)}_{uv} - R^{(2)}_{uv} \right| . \tag{4.3}$$

where the parameter β *(see* (4.2)*) is implicitly defined. In the rest of the chapter we work with* $\beta = 1$*, and dispense of the subscript* β *in* (4.3)*. In other words,*

$$\mathrm{RD} \stackrel{def}{=} \mathrm{RD}_1 . \tag{4.4}$$

Remark. *The metric given in Definition 4 can be used to compare graphs of two different sizes, by adding isolated vertices to both graphs until they have the same vertex set (this is why we must form the union* $V = V_1 \cup V_2$ *and compare the graphs over this vertex set). This method will give reasonable results when the overlap between* V_1 *and* V_2 *is large. In particular, if we are comparing graphs of size* n *and* $n+1$*, then we only need add one isolated vertex to the former so that we can compare it to the latter. This situation is illustrated in Figure 4.1.*

Remark. *An additional parameter* β *has been added to the definition. Changing* β *is equivalent to scaling the effective resistance before applying the function* $x \to x/(1+x)$*. Note that when* $\widehat{R} \ll \beta$*, then* $R \approx \widehat{R}/\beta$*, i.e. the renormalized resistance is approximately a rescaling of the effective*

resistance. Note that in this metric, two graphs are equal if they differ only in addition or removal of isolated vertices.

4.4 Graph Models

In our analysis, we will discuss two common random graph models, the classic model of Erdős and Rényi and the stochastic blockmodel which is a two-community extension of the same.

Definition 5 (Erdős-Rényi Random Graph)**.** *Let $n \in \mathbb{N}$ and let $p \in [0,1]$. We recall that the Erdős-Rényi random graph, $\mathcal{G}(n,p)$, is the probability space formed by the graphs defined on the set of vertices $[n]$, where edges are drawn randomly from $\binom{n}{2}$ independent Bernoulli random variables with probability p. In effect, the graph $G \in \mathcal{G}(n,p)$, with m edges, occurs with probability*

$$\text{Prob}\,(G) = p^m (1-p)^{\binom{n}{2}-m}. \tag{4.5}$$

We now introduce a model of a dynamic community network: the balanced, two-community stochastic blockmodel.

Definition 6 (Dynamic Stochastic Blockmodel)**.** *Let $n \in \mathbb{N}$, and let $p, q \in [0,1]$. We denote by $\mathcal{G}(n,p,q)$ the probability space formed by the graphs defined on the set of vertices $[n]$, constructed as follows.*

We split the vertices $[n]$ into two communities C_1 and C_2, formed by the odd and the even integers in $[n]$, respectively. We denote by n_1 and n_2 the size of C_1 and C_2, respectively. Edges within each community are drawn randomly from independent Bernoulli random variables with probability p. Edges between communities are drawn randomly from independent Bernoulli random variables with probability q. For $G \in \mathcal{G}(n,p,q)$, with m_1 and m_2 edges in communities C_1 and C_2, respectively, we have

$$\text{Prob}\,(G) = p^{m_1}(1-p)^{\binom{n_1}{2}-m_1} q^{m_2}(1-q)^{\binom{n_2}{2}-m_2}. \tag{4.6}$$

Remark. *In this work, we study nested sequences of random graphs, and will sometimes use the subscript n to denote the index of the corresponding element G_n in the graph process.*

Remark. *While our model assumes that the two communities have equal size, or differ at most by one vertex, the model can be extended to multiple communities of various sizes.*

We now introduce the notions of in-community, and cross-community degrees.

Definition 7. *For any vertex u, we denote by d_u the degree of u.*

*Furthermore, we denote by $d_u^{(in)}$ the number of **in-community** edges (u, v), where v is in the same community as u. Similarly, $d_u^{(out)}$ is the number of **cross-community** edges (u, v), where v is not in the same community as u.*

We denote by $\overline{d_n}$, the expected degree of the graph G_n. In the case where $G_n \in \mathcal{G}(n, p, q)$, then $\overline{d_n}$ refers to the expected in-community degree.

Finally, we denote by k_n the random number of cross-community edges between C_1 and C_2.

Remark. *In the case where $G_n \in \mathcal{G}(n, p, q)$, we have $n_1 = \lfloor (n+1)/2 \rfloor$, and $n_2 = \lfloor n/2 \rfloor$. Also, $\overline{d_n} = pn_1$ is the expected degree within community C_1, and $\overline{d_n} = pn_2$ is the expected degree within community C_2. Because asymptotically, $n_1 \sim n_2$, we ignore the dependency of $\overline{d_n}$ on the community when computing asymptotic behavior for large n. More precisely, we have the following results.*

Lemma 15. *Let $G_n \in \mathcal{G}(n, p, q)$, and let $\overline{d_{n1}} = pn_1$ be the expected degree within community C_1, and $\overline{d_{n2}} = pn_2$ be the expected degree within community C_2. We have*

(1) $\overline{d_{n1}} = \overline{d_{n2}} + \varepsilon p$, *where* $\varepsilon = 0$ *if* n *is even, or* $\varepsilon = 1$ *otherwise.*

(2) $\overline{d_{n1}}^2 = \overline{d_{n2}}^2 (1 + o(1))$.

(3) $\dfrac{1}{\overline{d_{n1}}^2} = \dfrac{1}{\overline{d_{n2}}^2} (1 + o(1))$.

(4) $\dfrac{1}{\overline{d_{n1}}} = \dfrac{1}{\overline{d_{n2}}} + \mathcal{O}\left(\dfrac{1}{\overline{d_n}^2}\right)$ *where* $\overline{d_n} = \overline{d_{n1}}$, *or* $\overline{d_n} = \overline{d_{n2}}$.

In the remainder of the text we write $1/\overline{d_n}$ *when either* $1/\overline{d_{n1}}$ *or* $1/\overline{d_{n2}}$ *could be used, and the error between the two terms is no larger than* $\mathcal{O}\left(1/\overline{d_n}^2\right)$.

Remark. *Although we use* $\mathcal{G}$ *for both random graph models, the presence of two or three parameters prevents ambiguity in our definitions. Note that* $\overline{d_n}$ *has different meanings when used in the setting of an Erdős-Rényi graph (where it denotes the expected degree) and a stochastic blockmodel (where it denotes the expected in-community degree). Also note that* $\overline{d_n}$ *is fixed for given* n *and* p *whereas* k_n *is a binomial random variable.*

4.5 Main Results

4.5.1 Informal Presentation of our Results

Before carefully stating the main result in the next subsection, we provide a back of the envelope analysis to help understand under what circumstances the resistance metric can detect an anomalous event in the dynamic growth of a stochastic blockmodel. In particular, we aim to detect whether cross-community edges are formed at a given timestep. In graphs with few cross-community edges, the addition of such an edge changes the geometry of the graph significantly. We will show that the creation of such edges can be detected with high probability when the average in-community degree dominates the number of cross-community edges.

Figure 4.2 illustrates the statement of the problem. As a new vertex (shown in magenta) is added to the graph G_n, the connectivity between the communities can increase, if edges are added between C_1 and C_2, or the communities can remain separated. If the addition of the new vertex promotes the merging of C_1 and C_2, then we consider the new graph G_{n+1} to be **structurally different** from G_n, otherwise G_{n+1} remains **structurally the same** as G_n (see Fig. 4.2).

The goal of the present work is to detect the fusion of the communities without identifying the communities. We show that the effective resistance yields a metric that is sensitive to changes in pattern of connections and connectivity structure between C_1 and C_2. Therefore it can be used to detect structural changes between G_n and G_{n+1} without detecting the structure present in G_n.

The informal derivation of our main result relies on the following three ingredients:

(1) each community in $\mathcal{G}(n,p,q)$ is approximately a "random graph" (Erdős-Rényi), $\mathcal{G}(n/2,p)$;

(2) the effective resistance between two vertices u,v within $\mathcal{G}(n/2,p)$ is concentrated around $2/\overline{d_n} = 2/(p(n/2-1)$;

(3) the effective resistance between $u \in C_1$ and $v \in C_2$ depends only on the bottleneck formed by the k_n cross-community edges, $\widehat{R}_{uv} \approx 1/k_n$.

We now proceed with an informal analysis of the changes in effective resistance distance when the new vertex, $n+1$, is added to the stochastic blockmodel G_n.

We first consider the "null hypothesis" where no cross-community edges is formed when vertex $n+1$ is added to the graph. All edges are thus created in the community of $n+1$, say C_1 (without any loss of generality). Roughly $pn/2$ new edges are created, and thus about $\mathcal{O}(n)$ vertices are affected by the addition of these new edges to C_1.

Because the effective resistance between any two vertices u,v in C_1 is now concentrated around $2/[p(n_1-1+1)] \geq 2/(\overline{d_n}+1)$, the changes in resistance after the addition of vertex $n+1$ is bounded by

$$\Delta\widehat{R}_{uv} \leq \frac{2}{\overline{d_n}} - \frac{2}{\overline{d_n}+1} = \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right). \tag{4.7}$$

Although, one would expect that only vertices in community C_1 (wherein $n+1$ has been added) be affected by this change in effective resistance, a more detailed analysis shows that vertices in C_2 slightly benefit of the increase in connectivity within C_1.

We now consider the alternate hypothesis, where at least one cross-community edge is formed after adding $n+1$ (see Fig. 4.2-bottom right). This additional cross-community edge has an effect on all pairwise effective resistances. Nevertheless, the most significant perturbation in $\widehat{R}_{uv}$ occurs for the $n/2 \times n/2$ pair of vertices in $C_1 \times C_2$. Indeed, if $u \in C_1$ and $v \in C_2$, the change in effective

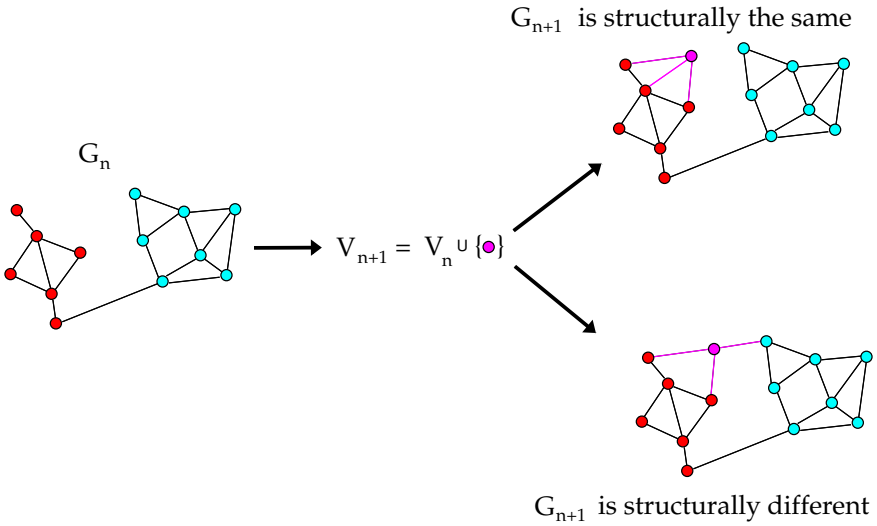

Figure 4.2: The dynamic stochastic blockmodel G_n is comprised of two communities (C_1: red and C_2: blue). As a new (magenta) vertex is added, the new graph G_{n+1} can remain structurally the same – if no new edges are created between C_1 and C_2 – or can become structurally different if the communities start to merge with the addition of new edges between C_1 and C_2.

resistance becomes

$$\Delta\widehat{R}_{uv} \approx \frac{1}{k_n} - \frac{1}{k_n + 1} = \mathcal{O}\left(\frac{1}{k_n^2}\right). \tag{4.8}$$

In summary, we observe asymptotic separation of the two regimes precisely when $k_n/\overline{d_n} \to 0$. We should therefore be able to use the renormalized resistance distance to test the null hypothesis that no edge is added between C_1 and C_2, and that G_n and G_{n+1} are structurally the same.

We will now introduce the main character of this work: the dynamic stochastic block model. We will then provide a precise statement of our result. In particular, we hope to elucidate our model of a dynamic community graph, in which at each time step a new vertex joins the graph and forms connections with previous vertices. The idea of graph growth as a generative mechanism is commonplace for models such as preferential attachment, but is less often seen for models such as Erdős-Rényi and the stochastic blockmodel.

4.5.2 The Dynamic Stochastic Blockmodel

Our growth mechanism adds a vertex to a stochastic blockmodel of size $n-1$, assigns it to a community according to the parity of n, and then connects it to each member of its community with probability p and each member of the opposite community with probability q. Continuing in this way generates a sequence of graphs $\{G_n\}_{n=1}^{\infty}$. If $n < N$ for some $N > 0$, then each G_n is the subgraph of G_N induced by $\{1, 2, \ldots, n\}$. It is not hard to see that any element of this sequence will, independent of the others, be distributed as in Definition 6. Note that here, the parameters $\{p, q\}$ are fixed. We interpret this sequence as a time-sampled growing graph.

If we are only interested in this infinite sequence up to a time N, then we could just as well form the graph $G \sim \mathcal{G}(N, p, q)$ and generate induced subgraphs from this graph. In practice, this is what we do. Since we look at the time step n to $n+1$, our result is phrased in terms of subgraphs of G_{n+1}.

In the above paragraph, the parameters p and q are fixed. However, our asymptotic result contains an index p_n. This can be understood as follows. We fix an $n > 0$, and corresponding

Graph Sequence					Parameters
$\underbrace{G_1}_{D_1}$	G_2	G_3	G_4	...	$\{p_1, q_1\}$
G_1	$\underbrace{G_2}_{D_2}$	G_3	G_4	...	$\{p_2, q_2\}$
G_1	G_2	$\underbrace{G_3}_{D_3}$	G_4	...	$\{p_3, q_3\}$
$\vdots$					$\vdots$

Figure 4.3: Diagram showing that we generate a sequence of graphs for each set of parameters $\{p_n, q_n\}$. We then measure the renormalized resistance distance $\mathrm{RD}(G_n, G_{n+1})$ between time steps n and $n+1$.

p_n and q_n. We then generate a stochastic blockmodel $G \sim \mathcal{G}(N, p_n^{(\mathrm{in})}, p_n^{(\mathrm{out})})$ for some $N \geq n+1$. We look at the renormalized resistance distance $\mathrm{RD}(G_n, G_{n+1})$ between the dynamic graph at two adjacent time steps n and $n+1$[3] . Said another way, we look at the distance between the subgraphs G_n and G_{n+1}. Figure 4.3 gives a diagrammatic explanation. S

Such a model may seem strange, but is in fact quite realistic. Consider a large social network. We are interested in examining small fluctuations (a single member joining the network) over fast timescales (days or hours). We would expect that the connectivity patterns of members of the network will change over much longer timescales (years), and thus can treat the parameters of the model as fixed when analyzing the network and performing anomaly detection.

In the stochastic blockmodel, each vertex u belongs to a community within the graph. If vertex u forms no cross-community edges, then the geometry of the graph is structurally the same. However, if u forms at least one cross-community edge, then (depending on the geometry of the preceding graphs in the sequence) the geometry may change significantly. We examine in what regimes of p_n and q_n we can differentiate between the two situations with high probability.

We phrase the result in terms of a hypothesis test, with the null hypothesis being that no cross-community edges have been formed in step $n+1$.

[3] We might be more careful, and assign each sequence two indices, n corresponding to the parameter and m corresponding to the subgraph. However, we find this notation to be unnecessarily cumbersome, so we omit it.

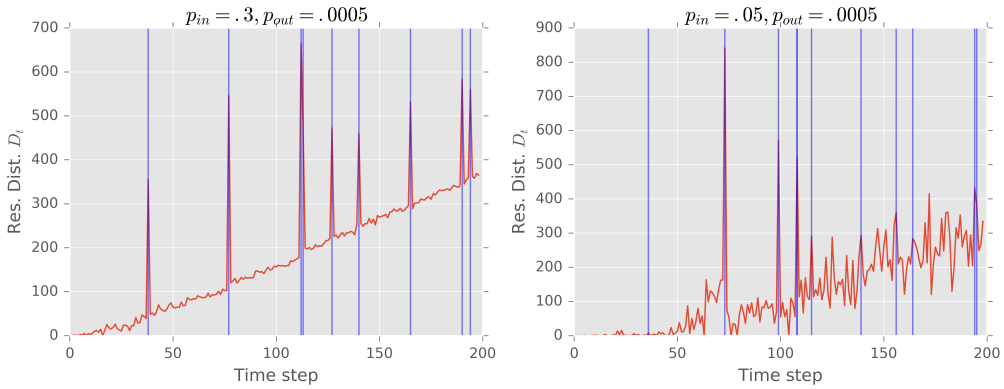

Figure 4.4: A typical time series of D_n for a dynamic stochastic blockmodel. The red curve is the distance between time steps D_n and the blue vertical lines mark the formation of cross-community connections. Two different regimes are compared. On the left, the formation of cross-community edges is easily discernible, while on the right, such an event is quickly lost in the noise.

Figure 4.4 shows an example of such a time series for a dynamic stochastic blockmodel. We see that when the in-community connectivity is much greater than the cross community connectivity, the formation of cross-community edges is easily discernible (left figure). However, when the level of connectivity is insufficiently separated, then the formation of cross-community edges is quickly lost in the noise. Our result clarifies exactly what is meant when we say that the parameters p_n and q_n are "well separated." Our main result is given by the following theorem.

Theorem 1. *Let* $G_{n+1} \sim \mathcal{G}(n+1, p_n, q_n)$ *be a stochastic blockmodel with* $p_n = \omega(\log n/n)$, $q_n = \omega\left(1/n^2\right)$, $q_n = o(p_n/n)$, *and* $p_n = \mathcal{O}(1/\sqrt{n})$. *Let* G_n *be the subgraph induced by the vertex set* $[n]$, *with* m_n *edges. Let* $D_n = \mathrm{RD}(G_n, G_{n+1})$ *be the normalized effective resistance distance,* RD, *defined in* (4.3).

To test the hypothesis

$$H_0 : \quad k_n = k_{n+1} \tag{4.9}$$

versus

$$H_1 : \quad k_n < k_{n+1} \tag{4.10}$$

we use the test based on the statistic Z_n defined by

$$Z_n \stackrel{def}{=} \frac{16m_n^2}{n^4}\left(D_n - n\right), \tag{4.11}$$

where we accept H_0 if $Z_n < z_\alpha$ and accept H_1 otherwise. Then, for every $0 < \alpha \leq 1$, there exists a threshold z_α for the rejection region satisfies

$$\mathrm{Prob}_{H_0}\left(Z_n \geq z_\alpha\right) \leq \alpha \quad as \quad n \to \infty, \tag{4.12}$$

and

$$\mathrm{Prob}_{H_1}\left(Z_n \geq z_\alpha\right) \to 1 \quad as \quad n \to \infty. \tag{4.13}$$

The test has therefore asymptotic level α and asymptotic power 1.

In practice, it would be desirable to have an analytical expression for the constant $C(\varepsilon)$ such that we can compute a level ε test,

$$\mathrm{Prob}\left(Z_n \geq C(\varepsilon)\right) \leq \varepsilon \quad \text{under the null hypothesis.}$$

Unfortunately, our technique of proof, which is based on the asymptotic behavior of $\widetilde{D}_n$ does not yield such a constant. A more involved analysis, based on finite sample estimates of the distance, would be needed, and would yield an important extension of the present work. The results shown in Figure 4.5 suggest that one could numerically estimate a $1-\varepsilon$ point wise confidence interval for $\widetilde{D}_n$ with a bootstrapping technique; the details of such a construction are the subject of ongoing investigation.

4.6 Experimental Analysis of Dynamic Community Networks

Figure 4.5 shows numerical evidence supporting Theorem 1. The empirical distribution of Z_n is computed under the null hypothesis (solid line) and the alternate hypothesis (dashed line). The data are scaled so that the empirical distribution in the case where the null hypothesis is true has zero mean and unit variance. In the left and right figures, the density of edges remains the same within each community, $p_n = \log^2 n/n$.

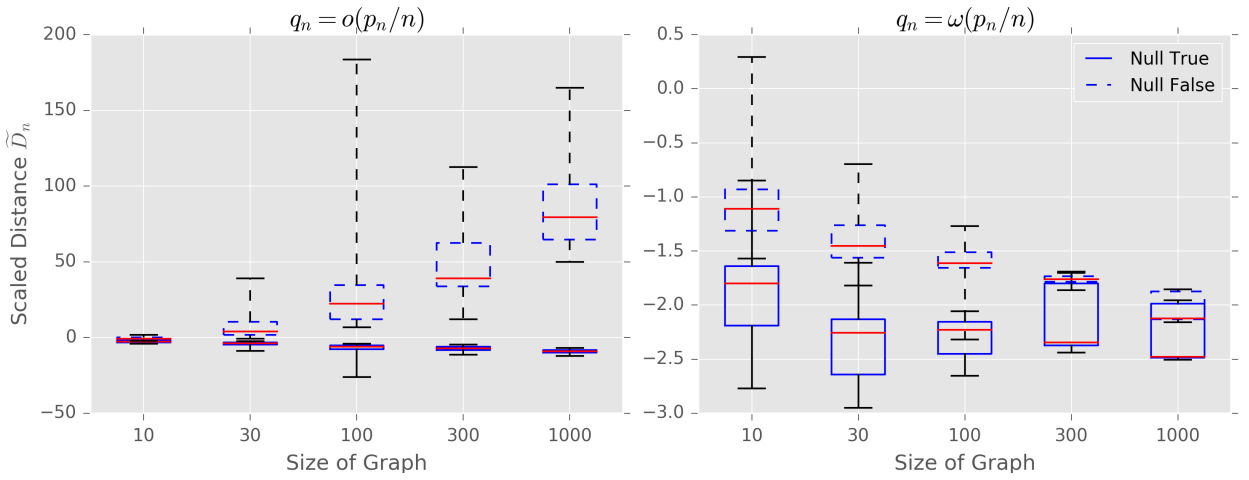

Figure 4.5: Empirical distribution of distance Z_n as defined in (4.11) showing separation of distributions when $q_n/p_n = o(1/n)$ (left) and overlap when $q_n/p_n = \omega(1/n)$ (right). The box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend from the box to show the full range of the data. Note that the y-axis is logarithmic in the left figure and linear in the right.

The plot on the left of Fig. 4.5 illustrates a case where the density of cross-community edges remains sufficiently low – $q_n = \log n/n^2$ – and the test statistic can detect the creation of novel cross-community edges (alternate hypothesis) without the knowledge of k_n, or the identification of the communities.

On the right, the density of cross-community edges is too large ($q_n = \log^2 n/n^{3/2}$) for the statistic to be able to detect the creation of novel cross-community edges. In that case the hypotheses of Theorem 1 are no longer satisfied.

In addition to the separation of distributions guaranteed by Theorem 1 when $q_n/p_n = o(1/n)$, we observe a separation between the two distributions when $q_n/p_n = \Theta(1/n)$ (not shown), suggesting that the hypotheses of Theorem 1 are optimal. In the regime where $q_n = \omega(p_n/n)$, shown in Fig. 4.5-right, the two distributions overlap.

4.7 Discussion

At first glance, our result may seem restrictive compared to existing results regarding community detection in the stochastic blockmodel. However, such a comparison is ill-advised, as we do not propose this scheme as a method for community detection. For example, Abbe et al. have shown that communities can be recovered asymptotically almost surely when $p_n = \alpha \log(n)/n$ and $q_n = \beta \log(n)/n$, given that $(\alpha + \beta)/2 - \sqrt{\alpha\beta} > 1$ [2]. Their method uses an algorithm that is designed specifically for the purpose of community detection, whereas our work provides a very general tool, which can be applied on a broad range of dynamic graphs, albeit without the theoretical guarantees that we derive for the dynamic stochastic block model.

Furthermore, the "efficient" algorithm proposed by Abbe et al. is only proven to be polynomial time, whereas resistance matrices can be computed in near-linear or quadratic time, for the approximate [136] and exact effective resistance respectively. This allows our tool to be of immediate practical use, whereas results such as those found in [2] are of a more theoretical flavor.

Some argue against the use of the effective resistance to analyze connectivity properties of a

graph. In [151] it is shown that

$$\left|\widehat{R}_{uv} - \frac{1}{2}\left(\frac{1}{d_u} + \frac{1}{d_v}\right)\right| \leq \left(\frac{1}{1-\lambda_2} + 2\right)\frac{w_{max}}{\delta_n^2}, \tag{4.14}$$

where λ_2 is the second largest eigenvalue of the normalized graph Laplacian, d_v is the degree of vertex v, δ_n is the minimum degree, and w_{max} is the maximum edge weight. If the right-hand side of (4.14) converges to 0, then the effective resistance will converge to the average inverse degree of u and v.

Luxburg et al. argue that the result (4.14) implies that when the bound converges to zero, the resistance will be uninformative, since it depends on local properties of the vertices and not global properties of the graph. Fortunately, this convergence can coexist peacefully alongside our result. In particular, if both expected degree and minimum degree approach infinity with high probability, as they will when both q_n and p_n are $\omega\left(1/n^2\right)$, then such convergence will itself occur with high probability (see (4.28) for the relevant spectral gap bound). Since we only care about relative changes in resistance between G_n and G_{n+1} in cases where cross-community edges are and are not formed, this is no problem for us. That said, the warning put forth by Luxburg et al. is well taken; we must be careful to make sure that we understand the expected behavior of the distance $\mathrm{RD}(G_n, G_{n+1})$ and compare the observed behavior to this expected behavior rather than evaluate it on an absolute scale, since in many situations of interest this distance will converge to zero as the graph grows.

Luxburg et al. have pointed out that the resistance can be fickle when used on graphs with high connectivity, which is to say a small spectral bound $(1-\lambda_2)^{-1}$. We now know in which circumstances this will become an issue when looking at simple community structure. Further investigation is needed to know when other random graph models such as the small-world or preferential attachment model will be susceptible to analysis via the renormalized resistance metric. We are also curious to apply this analysis to a variety of real-world data sets to examine the efficacy of our approach in such complicated contexts.

4.8 Appendix

4.8.1 Asymptotic Notations

If $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ are infinite sequences. The notations on the left have the interpretation on the right,

$$\begin{aligned}
a_n &= \mathcal{O}(b_n) && \exists n_0 > 0, \exists c > 0, \forall n \geq n_0, && 0 \leq |a_n| \leq c|b_n|, \\
a_n &= o(b_n) && \forall c > 0, \exists n_0 \geq 0, \forall n \geq n_0, && 0 \leq |a_n| \leq c|b_n|, \\
a_n &= \omega(b_n) && \forall c > 0, \exists n_0 \geq 0, \forall n \geq n_0, && 0 \leq c|b_n| < |a_n|, \\
a_n &= \Theta(b_n) && \exists c_1, c_2 > 0, \exists n_0 \geq 0, \forall n \geq n_0, && 0 \leq c_1|b_n| \leq |a_n| \leq c_2|b_n|.
\end{aligned}$$

We can adapt any of the above statements to doubly-indexed sequences $a_{n,k}$ and $b_{n,k}$ by requiring that there exist an $n_0 \geq 0$ such that the conditions on the right hold for all $n, k \geq n_0$.

4.8.2 Notations Used in the Chapter

4.8.3 Proof of Main Result

We begin by proving a lemma that allows us to transfer bounds on changes in effective resistance into bounds on changes in renormalized resistances.

Lemma 16. *Suppose that $\widehat{R}_1$ and $\widehat{R}_2$ are two effective resistances. If*

$$C_1 \leq \left|\widehat{R}_1 - \widehat{R}_2\right| \leq C_2,$$

*then the corresponding **renormalized** resistances obey*

$$\frac{C_1}{(\widehat{R}_1 + 1)(\widehat{R}_2 + 1)} \leq |R_1 - R_2| \leq C_2 \tag{4.15}$$

Proof. Recall that the renormalized resistance corresponding to $\widehat{R}$ is given by $R = f(\widehat{R})$ where $f(x) = x/(x+1)$. The mean value theorem thus implies that

$$|R_1 - R_1| \leq \sup_{x \in \mathbb{R}} |f'(x)| \left|\widehat{R}_1 - \widehat{R}_2\right| \leq \left|\widehat{R}_1 - \widehat{R}_2\right| \leq C_2.$$

$[n]$	The set of natural numbers $\{1, \ldots, n\}$.
$\mathcal{G}(n, p)$	Erdős-Rényi random graph with parameters n and p
$\mathcal{G}(n, p, q)$	stochastic blockmodel with parameters n, p, and q
G_n	Subgraph of a graph G induced by vertex set $[n]$
$\widehat{R}_{uv}$	Effective resistance between u and v
R_{uv}	Renormalized effective resistance between u and v
$R_{uv}^{(n)}$	Renormalized effective resistance between u and v in G_n
d_u	Degree of vertex u (random variable)
$d_u^{(out)}$	cross-community degree of vertex u (random variable)
$\overline{d_n}$	Mean degree (Erdős-Rényi) or expected in-community degree (stochastic blockmodel)
k_n	Number of cross-community edges (random variable)
$\overline{k_n}$	Mean number of cross-community edges
m_n	Total number of edges (random variable)
$\overline{m}_n$	Mean number of edges
$\mathrm{RP}(\cdot, \cdot)$	Resistance-perturbation distance
$\mathrm{RD}(\cdot, \cdot)$	Renormalized resistance distance (with $\beta = 1$)
$a_n = \mathcal{O}(b_n)$	a_n is asymptotically bounded by b_n
$a_n = \omega(b_n)$	a_n asymptotically dominates b_n
$a_n = o(b_n)$	a_n is asymptotically dominated by b_n

Table 4.1: Table of commonly used notation.

To obtain the lower bound, we compute:

$$|R_1 - R_1| = \left| \frac{\widehat{R}_1}{\widehat{R}_1 + 1} - \frac{\widehat{R}_2}{\widehat{R}_2 + 1} \right| = \frac{|\widehat{R}_1 - \widehat{R}_2|}{(\widehat{R}_1 + 1)(\widehat{R}_2 + 1)} \geq \frac{C_1}{(\widehat{R}_1 + 1)(\widehat{R}_2 + 1)}. \tag{4.16}$$

We have used the positivity of $\widehat{R}_1$ and $\widehat{R}_2$ in our calculation above. □

4.8.3.1 Resistance Deviations in Erdős-Rényi

We begin by analyzing the perturbations of the distance $\mathrm{RD}(G_n, G_{n+1})$, defined by (4.4), when $G \sim \mathcal{G}(n+1, p_n)$ is an Erdős-Rényi random graph. Our ultimate goal is to understand a stochastic blockmodel, and we will leverage our subsequent understanding of the Erdős-Rényi model to help us in achieving this goal.

Lemma 17. *Let $G \sim \mathcal{G}(n, p_n)$ be fully connected, with $p_n = \omega(\log n/n)$. For any two vertices u,v in G, we have*

$$\left| \widehat{R}_{uv} - \left(\frac{1}{d_u} + \frac{1}{d_v} \right) \right| = \mathcal{O}\left(\frac{1}{\overline{d_n}^2} \right) \quad \textit{with high probability.} \tag{4.17}$$

Remark. *The authors in [151] derive a slightly weaker bound,*

$$\left| \widehat{R}_{uv} - \left(\frac{1}{d_u} + \frac{1}{d_v} \right) \right| = o\left(\frac{1}{\delta_n} \right). \tag{4.18}$$

We need the tighter factor $\mathcal{O}\left(1/\overline{d_n}^2\right)$; and thus we derive the bound (4.17) using one of the key results (Proposition 5) in [151].

Proof. Define $\boldsymbol{D}$ to be the diagonal matrix with entries $d_1, \ldots, d_n$. Since all degrees are positive, we denote by $\boldsymbol{D}^{-1/2}$ the diagonal matrix with entries $1/\sqrt{d_1}, \ldots, 1/\sqrt{d_n}$. Let $\boldsymbol{A}$ be the adjacency matrix of G. Define $\boldsymbol{B} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, with eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. As explained above, we use Proposition 5 in [151] to bound the deviation of $\widehat{R}_{uv}$ away from $1/d_u + 1/d_v$,

$$\left| \widehat{R}_{uv} - \left(\frac{1}{d_u} + \frac{1}{d_v} \right) \right| \leq \frac{2}{\delta_n^2} \left(2 + \frac{1}{1 - \lambda_2} \right), \tag{4.19}$$

where δ_n is the minimum degree. Define

$$\varepsilon_n = \sqrt{6 \log(n) \overline{d_n}}. \tag{4.20}$$

We apply Chernoff's bound on the degree distribution,

$$\text{Prob}\left(\left|d_v - \overline{d_n}\right| \geq \varepsilon_n\right) \leq 2\exp\left(-\frac{\varepsilon_n^2}{3\overline{d_n}}\right).$$

Now,

$$\frac{\varepsilon_n^2}{3\overline{d_n}} = \frac{6\overline{d_n}\log n}{3\overline{d_n}} = \log n^2, \tag{4.21}$$

and thus

$$\text{Prob}\left(\left|d_v - \overline{d_n}\right| \geq \varepsilon_n\right) \leq \frac{2}{n^2}.$$

In the end, applying a union bound on all n vertices yields

$$\text{Prob}\left(\forall v \in [n], \left|d_v - \overline{d_n}\right| \geq \varepsilon_n\right) \leq \frac{2}{n}. \tag{4.22}$$

We consider d_v in the interval $[\overline{d_n} - \varepsilon_n, \overline{d_n} + \varepsilon_n]$. The mean value theorem implies that there exists $\tilde{d} \in (\overline{d_n}, d_v)$, or $\tilde{d} \in (d_v, \overline{d_n})$, such that

$$\left|\frac{1}{d_v^2} - \frac{1}{\overline{d_n}^2}\right| = \frac{2}{\tilde{d}^3}\left|\overline{d_n} - d_v\right|. \tag{4.23}$$

Now,

$$\frac{2}{\tilde{d}^3}\left|\overline{d_n} - d\right| \leq \frac{2\varepsilon_n}{(\overline{d_n} - \varepsilon_n)^3} = \frac{2\varepsilon_n}{\overline{d_n}^3}\frac{1}{\left(1 - \varepsilon_n/\overline{d_n}\right)^3}. \tag{4.24}$$

At last, we use the following elementary fact

$$0 \leq \frac{1}{(1-x)^3} \leq 1 + 12x, \quad \text{if } x < 1/4, \tag{4.25}$$

to conclude that

$$\forall v \in [n], \quad \left|\frac{1}{d_v^2} - \frac{1}{\overline{d_n}^2}\right| \leq \frac{2\varepsilon_n}{\overline{d_n}^3}\left(1 + 12\frac{\varepsilon_n}{\overline{d_n}}\right) \quad \text{with probability greater than } 2/n. \tag{4.26}$$

This eventually yields an upper bound on the inverse of the minimum degree squared,

$$\frac{1}{\delta_n^2} \leq \frac{1}{\overline{d_n}^2} + \frac{2\varepsilon_n}{\overline{d_n}^3}\left(1 + 12\frac{\varepsilon_n}{\overline{d_n}}\right) \quad \text{with probability greater than } 2/n. \tag{4.27}$$

To complete the proof of the lemma, we use a lower bound on the spectral gap $1-\lambda_2$. Because the density of edges is only growing faster than $\log n/n$, we use the optimal bounds given by [37]. Applied to the eigenvalue λ_2 of $\boldsymbol{B}$, Theorem 1.2 of [37] implies that

Theorem 2 ([37]). *If $\overline{d_n} > c \log n/n$, then with high probability,*

$$1 - \frac{c}{\sqrt{\overline{d_n}}} \leq 1 - \lambda_2 \leq 1 + \frac{c}{\sqrt{\overline{d_n}}}. \tag{4.28}$$

The lower bound in (4.28) yields the following upper bound, with high probability,

$$\frac{1}{1-\lambda_2} \leq 1 + \frac{c}{\sqrt{\overline{d_n}}}. \tag{4.29}$$

Using the bounds given by (4.27) with (4.29), which happen both with high probability, in (5.4) yields the advertised result. □

An important corollary of lemma 17 is the concentration of $\widehat{R}_{uv}$ around $2/\overline{d_n}$ (see also [151] for similar results),

Corollary 18. *Let $G \sim \mathcal{G}(n, p_n)$ be fully connected, with $p_n = \omega(\log n/n)$. With high probability,*

$$\left| \widehat{R}_{uv} - \frac{2}{\overline{d_n}} \right| = \frac{16}{\overline{d_n}^2} + o\left(\frac{1}{\overline{d_n}^2} \right). \tag{4.30}$$

The lemma provides a confidence interval for the effective resistance $\widehat{R}_{uv}$ centered at $d_u^{-1} + d_v^{-1}$, for pairs of vertices present in the graph G at time n. We now use this result to bound the change in (renormalized) resistance between a pair of vertices u and v present in G_n, when the graph grows from G_n to G_{n+1}.

Theorem 3. *Let $G_{n+1} \sim \mathcal{G}(n+1, p_n)$ be an Erdős-Rényi random graph with $p_n = \omega(\log n/n)$. Let G_n be the subgraph induced by the vertices $[n]$ in G_{n+1}, and let $\overline{d_n} = (n-1)p_n$ be the expected degree in G_n.*

The change in renormalized effective resistance, when the graph G_n becomes G_{n+1}, is given by

$$\max_{u<v\leq n} \left| R_{u,v}^{(n+1)} - R_{u,v}^{(n)} \right| = \mathcal{O}\left(\frac{1}{\overline{d_n}^2} \right) \quad \textit{with high probability.} \tag{4.31}$$

Remark. *It is important to note that the bound on changes in $R_{u,v}$ from time n to $n+1$ only holds for the nodes $u, v \in [n]$ that are already present in $\mathcal{G}(n, p_n)$. Indeed, for the new node $n+1$ that is added at time $n+1$, we have $\widehat{R}^{(n)}_{u,n+1} = \infty$, and thus $R^{(n)}_{u,n+1} = 1$. In this case, the bound in* (4.31) *is replaced by*

$$\left|R^{(n+1)}_{u,n+1} - R^{(n)}_{u,n+1}\right| = R^{(n)}_{u,n+1} - R^{(n+1)}_{u,n+1} = 1 - R^{(n+1)}_{u,n+1} \leq 1 \quad \textit{for any vertex } u \in [n]. \tag{4.32}$$

Proof. By Lemma 16, it suffices to prove the inequality with respect to the effective resistance, rather than the renormalized resistance.

Let $d^{(n)}_u$ denote the degree of vertex u in G_n, and similarly define $d^{(n+1)}_u$. Using the triangle inequality,

$$\begin{aligned}\left|\widehat{R}^{(n+1)}_{u,v} - \widehat{R}^{(n)}_{u,v}\right| \leq \left|\widehat{R}^{(n)}_{u,v} - \left(\frac{1}{d^{(n)}_u} + \frac{1}{d^{(n)}_v}\right)\right| + \left|\left(\frac{1}{d^{(n)}_u} + \frac{1}{d^{(n)}_v}\right) - \left(\frac{1}{d^{(n+1)}_u} + \frac{1}{d^{(n+1)}_v}\right)\right| \\ + \left|\widehat{R}^{(n+1)}_{u,v} - \left(\frac{1}{d^{(n+1)}_u} + \frac{1}{d^{(n+1)}_v}\right)\right|.\end{aligned} \tag{4.33}$$

From lemma 17, we obtain bounds on the first and third terms, of order $\mathcal{O}\left(1/\delta_n^2\right)$. Since $|d^{(n+1)}_u - d^{(n)}_u| \leq 1$, the middle term is of order $\mathcal{O}\left(1/\delta_n^2\right)$, which can in turn be bounded by a term of order $\mathcal{O}\left(1/\overline{d_n}^2\right)$ with high probability using 4.27. Putting everything together, we get

$$\left|\widehat{R}^{(n+1)}_{u,v} - \widehat{R}^{(n)}_{u,v}\right| = \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right).$$

The inequality is proven for the effective resistance, and using lemma 16 it also holds for the renormalized resistance. □

4.8.3.2 Effective resistances in the stochastic blockmodel

We now translate our understanding of the Erdős-Rényi random graph to the analysis of the stochastic blockmodel. As explained in lemma 15, in the following we write $1/\overline{d_n}$ when either $1/\overline{d_n}_1$ or $1/\overline{d_n}_2$ could be used, and the error between the two terms is no larger than $\mathcal{O}\left(1/\overline{d_n}^2\right)$.

We first recall that the number of cross-community edges, k_n, is a binomial distribution, and thus concentrates around its expectation $\mathbb{E}[k_n]$ for large n.

Lemma 19. *Let $G_n \sim \mathcal{G}(n, p_n, q_n)$ be a (balanced two-community) stochastic blockmodel with $p_n = \omega(\log n/n)$ and $q_n = \omega\left(1/n^2\right)$. There exists n_0, such that*

$$\forall n \geq n_0, \quad \frac{3}{4} < \frac{\mathbb{E}[k_n]}{k_n} < \frac{3}{2}, \quad \textit{with probability ¿ 0.9.} \tag{4.34}$$

Proof. The random variable k_n is binomial $B(n_1 n_2, q_n)$, where $n_1 = \lfloor (n+1)/2 \rfloor$, and $n_2 = \lfloor n/2 \rfloor$. We have $\mathbb{E}[k_n] = n_1 n_2 q_n$. We apply a Chernoff's bound on k_n to get

$$\text{Prob}\left(|k_n - \mathbb{E}[k_n]| > \varepsilon\right) < 2e^{-\varepsilon^2/(3\mathbb{E}[k_n])}. \tag{4.35}$$

Using $\varepsilon = 3\sqrt{\mathbb{E}[k_n]}$, we get

$$\text{Prob}\left(|k_n - \mathbb{E}[k_n]| > 3\sqrt{\mathbb{E}[k_n]}\right) < 2e^{-3} < 0.1 \tag{4.36}$$

or

$$1 - 3\frac{1}{\sqrt{\mathbb{E}[k_n]}} < \frac{k_n}{\mathbb{E}[k_n]} < 1 + 3\frac{1}{\sqrt{\mathbb{E}[k_n]}} \quad \text{with probability ¿ 0.9.} \tag{4.37}$$

Now, $\mathbb{E}[k_n] = q_n n_1 n_2 = q_n \mathcal{O}\left(n^2\right) = \omega(1)$, and thus $\lim_{n\to\infty} \mathbb{E}[k_n] = \infty$. Consequently $\exists n_0$ such that

$$\forall n \geq n_0, \quad \mathbb{E}[k_n] > 81, \tag{4.38}$$

and thus

$$\forall n \geq n_0, \quad \frac{3}{4} < \frac{\mathbb{E}[k_n]}{k_n} < \frac{3}{2}, \quad \text{with probability ¿ 0.9.} \tag{4.39}$$

□

Lemma 20 (Cross-community resistance bounds)**.** *Let $G_n \sim \mathcal{G}(n, p_n, q_n)$ be a (balanced two-community) stochastic blockmodel with $p_n = \omega(\log n/n)$ and $q_n = \omega\left(1/n^2\right)$. Let u and v be vertices in the communities C_1 and C_2, respectively. Let $\overline{d_n}$ be the expected in-community degree of C_1. Let k_n be the (random) number of cross-community edges. With high probability, the effective resistance $\widehat{R}_{uv}$ is bounded according to*

$$\frac{1}{k_n} \leq \widehat{R}_{u,v} \leq \frac{1}{k_n} + \frac{4}{\overline{d_n}} + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right). \tag{4.40}$$

Remark. *We recall (see lemma 15) that when we write $\overline{d_n}$, in 4.46, it either means the expected degree of C_1 or C_2.*

Remark. *The requirement that $p_n = \omega(\log(n)/n)$ guarantees that we are in a regime where resistances in the Erdős-Rényi graph converge to $2/\overline{d_n}$. The requirement that $q_n = \omega(1/n^2)$ guarantees that $\mathbb{E}[k_n] \to \infty$, and because of lemma 19, $k_n \to \infty$ with high probability. Finally, $q_n = o(p_n/n)$ guarantees that $\mathbb{E}[k_n] = o\left(\overline{d_n}\right)$, and using lemma 19 we have $k_n = o\left(\overline{d_n}\right)$ with high probability.*

Proof of the lemma 20. Without loss of generality, we assume that $u \in C_1$, and $v \in C_2$. To obtain the lower bound on $\widehat{R}_{uv}$ we use the Nash-Williams inequality [96], which we briefly recall here. Let u and v be two distinct vertices. A set of edges E_c is an **edge-cutset** separating u and u if every path from u to v includes an edge in E_c.

Lemma 21 (Nash-Williams, [96])**.** *If u and v are separated by K disjoint edge-cutsets $E_k, k = 1, \ldots, K$, then*

$$\sum_{k=1}^{K} \left[\sum_{(v_n, v_m) \in E_k} R_{n,m}^{-1} \right]^{-1} \leq \widehat{R}_{uv}, \qquad \text{where } (v_n, v_m) \text{ is an edge in the cutset } E_k. \tag{4.41}$$

Since the set of cross-community edges is a cutset for all pairs of vertices u and v in separate communities, and the size of this set is precisely k_n, we immediately obtain the desired lower bound.

The upper bound is obtained using the characterization of the effective resistance based on Thomson principle [94], which we recall briefly in the following. Let f be a flow along the edges E from u to v, and let

$$\mathcal{E}(f) = \sum_{e \in E} f^2(e) R_e, \tag{4.42}$$

be the energy of the flow f, where each undirected edge e in the sum is only counted once. A unit flow has strength one,

$$\operatorname{div}(f)(u) = -\operatorname{div}(f)(v) = 1. \tag{4.43}$$

Thomson's principle provides the following characterization of the effective resistance $\widehat{R}_{uw}$,

$$\widehat{R}_{uv} = \min \left\{ \mathcal{E}(f), f \text{ is a unit flow from } u \text{ to } v \right\}. \tag{4.44}$$

We use Thomson's principle in the following way: we construct a unit flow f from u to v. For this flow, the energy $\mathcal{E}(f)$ yields an upper bound on $\widehat{R}_{uw}$.

First, consider the case where neither u nor v are incident with any of the k_n cross-community edges, $e_i = (u_i, v_i), i = 1, \ldots, k_n$; where $u_i \in C_1$ and $v_i \in C_2$. Denote by f_i^u the unit flow associated with the effective resistance between u and u_i when only the edges in C_1 are considered. Similarly define f_i^v to be the unit flow associated with the effective resistance between v and v_i when when only the edges in C_2 are considered. Using the corollary 18, given by (4.30), we have with high probability,

$$\mathcal{E}(f_i^u) = \frac{2}{\overline{d_n}} + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right), \quad \text{and} \quad \mathcal{E}(f_i^v) = \frac{2}{\overline{d_n}} + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right). \tag{4.45}$$

We note that the expression of $\mathcal{E}(f_i^v)$ should involve the expected degree in C_2. As explained in lemma 15, we can use $\overline{d_n}$ since the difference between the two terms is absorbed in the $\mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right)$ term. Finally, let f_i^e be the flow that is 1 on edge e_i and 0 elsewhere. To conclude, we assemble the three flows and define

$$f(e) = \frac{1}{k_n} \sum_{i=1}^{k} \left\{ f_i^u(e) + f_i^e(e) + f_i^v(e) \right\},$$

which is a unit flow from u to v. Since $\mathcal{E}$ is a convex function, we can bound the energy of f via

$$\begin{aligned}\mathcal{E}(f) &= \mathcal{E}\left(\frac{1}{k_n}\sum_{i=1}^{k}\left\{f_i^u(e) + f_i^e(e) + f_i^v(e)\right\}\right) \leq \mathcal{E}\left(\frac{1}{k_n}\sum_{i=1}^{k} f_i^u(e)\right) + \mathcal{E}\left(\frac{1}{k_n}\sum_{i=1}^{k} f_i^e(e)\right) + \mathcal{E}\left(\frac{1}{k_n}\sum_{i=1}^{k} f_i^v(e)\right) \\ &\leq \frac{1}{k_n}\sum_{i=1}^{k}\mathcal{E}(f_i^u) + \frac{1}{k_n} + \frac{1}{k_n}\sum_{i=1}^{k}\mathcal{E}(f_i^v) = \frac{4}{\overline{d_n}} + \frac{1}{k_n} + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right).\end{aligned}$$

The final line holds with high probability. Note that we calculate the energy of the flow in the center term directly, whereas convexity is used to estimate the energy in the first and third term.

This upper bound also holds when the cross-community degree of u or v is nonzero. In this case, $u = u_i$ for some i. For this i, we can formally define the flow f_i^u between u and u_i to be the zero flow, which minimizes the energy trivially and has energy equal to the resistance between u and u_i (which is zero). Then the above calculation yields a smaller upper bound for the first and third terms. □

Remark. *Lemma 20 provides a first attempt at analysing the perturbation of the effective resistance under the addition of edges in the stochastic blockmodel. The upper bound provided by (4.40) is too loose to be useful, and we therefore resort to a different technique to get a tighter bound. The idea is to observe that the effective resistance is controlled by the bottleneck formed by the cross-community edges. We can get very tight estimate of the fluctuations in the effective resistance using a detailed analysis of the addition of a single cross-community edge. We use the Sherman–Morrison–Woodburry theorem [59] to compute a rank-one perturbation of the pseudo-inverse of the normalized graph Laplacian,* $\boldsymbol{L}^{\dagger}$. *The authors in [120] provide us with the exact expression that is needed for our work, see* (4.50). *The proof proceeds by induction on the number of cross-community edges,* k_n.

Theorem 4. *Let* $G_n \sim \mathcal{G}(n, p_n, q_n)$ *be a balanced, two community stochastic blockmodel with* $p_n = \omega(\log n/n)$, $q_n = \omega(1/n^2)$, *and* $q_n = o(p_n/n)$. *We assume that* G_n *is connected. The effective resistance between two vertices* u *and* v *is given by*

$$\widehat{R}_{u,v} = \frac{2}{\overline{d_n}} + \begin{cases} \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right), & \text{if } u \text{ and } v \text{ are in the same community,} \\ \frac{1}{k_n} + \frac{\alpha(k_n,u,v)}{\overline{d_n} k_n} + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right), & \text{otherwise.} \end{cases} \tag{4.46}$$

Also, conditioned on $k_n = k$ *the random variable* $\alpha(k_n, u, v)$ *is a deterministic function of* k, *and we have* $\alpha(k_n, u, v) = \mathcal{O}(k_n)$.

Proof. First, observe that Lemma 20 immediately implies that

$$-2k_n \leq \alpha(k_n, u, v) \leq 2k_n, \quad \forall u, v, \tag{4.47}$$

with high probability, so $\alpha = \mathcal{O}(k_n)$.

Next, let us show that the in-community resistances follow the prescribed form. The proof proceeds as follows: we derive the expression (4.46) conditioned on the random variable $k_n = k$, and we prove that $\alpha(k, u, v)$ is indeed a deterministic function in this case; the derivation of (4.46) is obtained by induction on k.

The engine of our induction is the update formula proven in [120]. This provides an exact formula (equation (4.50) below) for the change in resistance between any pair of vertices in a graph when a single edge is added or removed. The particular motivation of the authors in [120] is to calculate rank-1 updates to the pseudoinverse of the combinatorial graph Laplacian; however, it conspires that their formula is also very useful to inductively calculate resistances in the stochastic blockmodel.

We first consider the base case, where G_n is a balanced, two community stochastic blockmodel of size n with $k_n = 1$ cross-community edge. Denote this edge $e_1 = (u_1, v_1)$, where $u_1 \in C_1$ and $v_1 \in C_2$. We will refer to this graph as $G_n^{(1)}$.

The addition of a single edge connecting otherwise disconnected components does not change the resistance within those components, as it does not introduce any new paths between two vertices within the same component. We know that each community is an Erdős-Rényi graph with parameters p_n and $\lfloor n/2 \rfloor$ (or $\lceil n/2 \rceil$), so corollary 18 provides the expression for the effective resistance between two vertices within each community. A simple circuit argument allows us to obtain the resistance between u and v in separate communities via

$$\widehat{R}_{u,v} = \widehat{R}_{u,u_1} + \widehat{R}_{u_1,v_1} + \widehat{R}_{v_1,v}. \tag{4.48}$$

If $u \neq u_1$ and $v \neq v_1$, then we combine Nash-Williams and corollary 18 yields to get

$$\widehat{R}_{u,v} = \frac{1}{k} + \frac{4}{\overline{d_n}} + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right).$$

If $u = u_1$ and/or $v = v_1$ then the appropriate resistances are set to zero in (4.48).

In summary, for arbitrary pairs (u, v) in $G_n^{(1)}$, we have

$$\widehat{R}_{u,v} = \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha(u,v)}{k\overline{d_n}} + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right),$$

where

$$\alpha(u,v) = 2(2 - d_u^{(out)} - d_v^{(out)}). \tag{4.49}$$

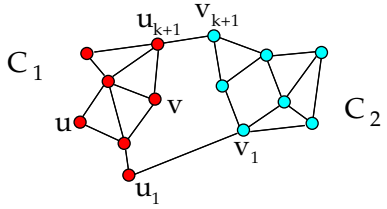

Figure 4.6: Balanced, two community stochastic blockmodel $G_n^{(k+1)}$ of size n with $k_n = k+1$ cross-community edges. Vertices u and v are in the same community C_1

This establishes the base case for (4.46). We now assume that (4.46) holds for any balanced, two community stochastic blockmodel of size n with $k_n = k$ cross-community edges. We consider a balanced, two community stochastic blockmodel $G_n^{(k+1)}$ of size n with $k_n = k+1$ cross-community edges. We denote the cross-community edges by $e_i = (u_i, v_i), i = 1, \ldots, k+1$, where $u_i \in C_1$ and $v_i \in C_2$.

Finally, we denote by $G_n^{(k)}$ the balanced, two community stochastic blockmodel with k cross-community edges obtained by removing the edge $e_{k+1} = (u_{k+1}, v_{k+1})$ from $G_n^{(k+1)}$.

Let $\widehat{R}$ denote the effective resistances in $G_n^{(k)}$ and $\widehat{R}'$ denote the effective resistances in $G_n^{(k+1)}$.

Since $G_n^{(k)}$ is obtained by removing an edge from $G_n^{(k+1)}$, we can apply equation (11) in [120] to express $\widehat{R}'$ from $\widehat{R}$,

$$\widehat{R}'_{u,v} = \widehat{R}_{u,v} - \frac{\left[(\widehat{R}_{u,v_{k+1}} - \widehat{R}_{u,u_{k+1}}) - (\widehat{R}_{v,v_{k+1}} - \widehat{R}_{v,u_{k+1}})\right]^2}{4(1 + \widehat{R}_{u_{k+1},v_{k+1}})}. \tag{4.50}$$

In the following we use the induction hypothesis to compute $\widehat{R}'$ using (4.50). We first consider the case where the vertices u and v belong to the same community, say C_1 without loss of generality (see Fig. 4.6).

We need to consider the following three possible scenarios:

(1) $u \neq u_{k+1}$ and $v \neq v_{k+1}$,

(2) $u = u_{k+1}$ and $v \neq v_{k+1}$,

(3) $u \neq u_{k+1}$ and $v = v_{k+1}$.

We will treat the first case; the last two cases are in fact equivalent, and are straightforward consequences of the first case. From the induction hypothesis we have

$$\widehat{R}_{u,v_{k+1}} = \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha(k,u,v_{k+1})}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right), \tag{4.51}$$

$$\widehat{R}_{u,u_{k+1}} = \frac{2}{\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right), \tag{4.52}$$

$$\widehat{R}_{v,v_{k+1}} = \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha(k,v,v_{k+1})}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right), \tag{4.53}$$

$$\widehat{R}_{v,u_{k+1}} = \frac{2}{\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right), \tag{4.54}$$

$$\widehat{R}_{u_{k+1},v_{k+1}} = \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha(k,u_{k+1},v_{k+1})}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right). \tag{4.55}$$

Substituting these expression into (4.50), we get

$$\widehat{R}'_{u,v} = \widehat{R}_{u,v} - \frac{\left[\left(\frac{1}{k} + \frac{\alpha(k,u,v_{k+1})}{k\overline{d_n}}\right) - \left(\frac{1}{k} + \frac{\alpha(k,v,v_{k+1})}{k\overline{d_n}}\right) + \mathcal{O}\left(1/\overline{d_n}^2\right)\right]^2}{4(1 + \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha(k,u_{k+1},v_{k+1})}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right))} \tag{4.56}$$

$$= \widehat{R}_{u,v} - \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right) \frac{\left[\frac{\alpha(k,u,v_{k+1}) - \alpha(k,v,v_{k+1})}{k} + \mathcal{O}\left(1/\overline{d_n}\right)\right]^2}{4\left(1 + \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha(k,u_{k+1},v_{k+1})}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)\right)} \tag{4.57}$$

Because α is bounded with high probability (see (4.47)), we have

$$\frac{\alpha(k,u,v_{k+1}) - \alpha(k,v,v_{k+1})}{k} + \mathcal{O}\left(1/\overline{d_n}\right) = \mathcal{O}(1) \quad \text{with high probability,} \tag{4.58}$$

and also

$$1 + \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha(k,u_{k+1},v_{k+1})}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right) = 1 + \frac{1}{k} + \mathcal{O}\left(1/\overline{d_n}\right) \quad \text{with high probability,} \tag{4.59}$$

which implies

$$\frac{\left[\frac{\alpha(k,u,v_{k+1}) - \alpha(k,v,v_{k+1})}{k} + \mathcal{O}\left(1/\overline{d_n}\right)\right]^2}{4\left(1 + \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha(k,u_{k+1},v_{k+1})}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)\right)} = \mathcal{O}(1) \quad \text{with high probability.} \tag{4.60}$$

We conclude that

$$\widehat{R}'_{u,v} = \widehat{R}_{u,v} + \mathcal{O}\left(1/\overline{d_n}^2\right) \quad \text{with high probability.} \tag{4.61}$$

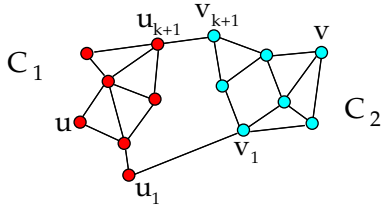

Figure 4.7: Balanced, two community stochastic blockmodel $G_n^{(k+1)}$ of size n with $k_n = k+1$ cross-community edges. The vertices u and v are in different communities, $u \in C_1$ and $v \in C_2$.

This completes the induction, and the proof of (4.46) in the case where u and v belong to the same community.

We now consider the case where $u \in C_1$ and $v \in C_2$ (see Fig. 4.7). As above, we need to consider the following three possible scenarios:

(1) $u \neq u_{k+1}$ and $v \neq v_{k+1}$,

(2) $u = u_{k+1}$ and $v \neq v_{k+1}$,

(3) $u \neq u_{k+1}$ and $v = v_{k+1}$.

Again, we only prove the first case; the last two equivalent cases are straightforward consequences of the first case. From the induction hypothesis we now have

$$\widehat{R}_{u,u_{k+1}} = \frac{2}{\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right), \tag{4.62}$$

$$\widehat{R}_{v,v_{k+1}} = \frac{2}{\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right), \tag{4.63}$$

$$\widehat{R}_{u,v_{k+1}} = \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha(k,u,v_{k+1})}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right), \tag{4.64}$$

$$\widehat{R}_{v,u_{k+1}} = \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha(k,v,v_{k+1})}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right), \tag{4.65}$$

$$\widehat{R}_{u_{k+1},v_{k+1}} = \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha(k,u_{k+1},v_{k+1})}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right). \tag{4.66}$$

To reduce notation clutter, we use some abbreviated notation to denote various α terms associated with the vertices of interest (see Fig. 4.8),

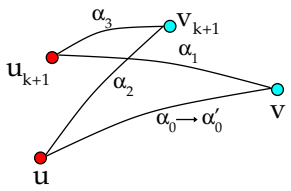

Figure 4.8: Coefficients $\alpha_0, \ldots, \alpha_3$ in (4.46) for several pairs of vertices in $G_n^{(k+1)}$.

$$\alpha_0 \stackrel{\text{def}}{=} \alpha(k_n^{(k)}, u, v),$$
$$\alpha_1 \stackrel{\text{def}}{=} \alpha(k_n^{(k)}, u_{k+1}, v),$$
$$\alpha_2 \stackrel{\text{def}}{=} \alpha(k_n^{(k)}, u, v_{k+1}),$$
$$\alpha_3 \stackrel{\text{def}}{=} \alpha(k_n^{(k)}, u_{k+1}, v_{k+1}).$$

Let us denote the decrease in effective resistance by $\Delta\widehat{R}$,

$$\Delta\widehat{R} \stackrel{\text{def}}{=} \widehat{R}_{u,v} - \widehat{R}'_{u,v}. \tag{4.67}$$

From (4.50), we have

$$\begin{aligned}
\Delta\widehat{R} &= \frac{\left[\left(\frac{1}{k} + \frac{\alpha_2}{k\overline{d_n}}\right) - \left(-\frac{1}{k} - \frac{\alpha_1}{k\overline{d_n}}\right) + \mathcal{O}\left(1/\overline{d_n}^2\right)\right]^2}{4\left(1 + \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha_3}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)\right)} = \frac{\left[\frac{2}{k}\left(1 + \frac{\alpha_1+\alpha_2}{2\overline{d_n}}\right) + \mathcal{O}\left(1/\overline{d_n}^2\right)\right]^2}{4\left(1 + \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha_3}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)\right)} \\
&= \frac{1}{k^2}\left[1 + \frac{\alpha_1+\alpha_2}{2\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)\right]^2 \frac{1}{1 + \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha_3}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)} \\
&= \left(\frac{1}{k^2} + \frac{\alpha_1+\alpha_2}{\overline{d_n}k^2} + \frac{(\alpha_1+\alpha_2)^2}{4(\overline{d_n}k)^2} + \mathcal{O}\left(1/\overline{d_n}^2\right)\right) \frac{1}{1 + \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha_3}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)} \\
&= \left(\frac{1}{k^2} + \frac{\alpha_1+\alpha_2}{\overline{d_n}k^2} + \mathcal{O}\left(1/\overline{d_n}^2\right)\right) \frac{1}{1 + \frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha_3}{k\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)}.
\end{aligned} \tag{4.68}$$

At this juncture, we need to expand $\left(1 + 1/k + 2/\overline{d_n} + \alpha_3/(k\overline{d_n}) + \mathcal{O}\left(1/\overline{d_n}^2\right)\right)^{-1}$ using a Taylor series, which is possible when $1/k + 2/\overline{d_n} + \alpha_3/(k\overline{d_n}) < 1$. Since we are interested in the

large n asymptotic, we can assume that for n sufficiently large $2/\overline{d_n} + \alpha_3/(k\overline{d_n}) < 1/2$, and thus we need to guarantee that $k \geq 2$.

The case $k = 1$ needs to be handled separately. Setting $k = 1$ into (4.68) yields

$$\Delta\widehat{R} = \frac{1 + \frac{\alpha_1+\alpha_2}{\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)}{2 + \frac{2+\alpha_3}{\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)} = \frac{1}{2}\frac{1 + \frac{\alpha_1+\alpha_2}{\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)}{1 + \frac{2+\alpha_3}{2\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)}, \tag{4.69}$$

for n sufficiently large $(2 + \alpha_3)/\overline{d_n} < 1$, thus

$$\begin{aligned}
\Delta\widehat{R} &= \frac{1}{2}\left(1 + \frac{\alpha_1+\alpha_2}{\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)\right)\left(1 - \frac{2+\alpha_3}{2\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)\right) \\
&= \frac{1}{2}\left(1 + \frac{2\alpha_1 + 2\alpha_2 - 2 - \alpha_3}{2\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right)\right) \\
&= \frac{1}{2} + \frac{\alpha_1 + \alpha_2 - 1 - \alpha_3/2}{2\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right),
\end{aligned} \tag{4.70}$$

which leads to

$$\begin{aligned}
\widehat{R}'_{uv} = \widehat{R}_{uv} - \Delta\widehat{R} &= 1 + \frac{2}{\overline{d_n}} + \frac{\alpha_0}{\overline{d_n}} - \frac{1}{2} - \frac{\alpha_1 + \alpha_2 - 1 - \alpha_3/2}{2\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right), \\
&= \frac{1}{2} + \frac{2}{\overline{d_n}} + \frac{2\alpha_0 - \alpha_1 - \alpha_2 + 1 + \alpha_3/2}{2\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right).
\end{aligned} \tag{4.71}$$

Let

$$\alpha'_0 \stackrel{\text{def}}{=} 2\alpha_0 - \alpha_1 - \alpha_2 + 1 + \alpha_3/2, \tag{4.72}$$

then we have

$$\widehat{R}'_{uv} = \frac{1}{2} + \frac{2}{\overline{d_n}} + \frac{\alpha'_0}{2\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right), \tag{4.73}$$

which matches the expression given in (4.46) for $k = 2$. This completes the induction for $k = 1$.

We now proceed to the general case where $k \geq 2$. In that case, we use a Taylor series expansion of $\left(1 + 1/k + 2/\overline{d_n} + \alpha_3/(k\overline{d_n}) + \mathcal{O}\left(1/\overline{d_n}^2\right)\right)^{-1}$, and we get

$$\left(1 + 1/k + 2/\overline{d_n} + \alpha_3/(k\overline{d_n}) + \mathcal{O}\left(1/\overline{d_n}^2\right)\right)^{-1} = \sum_{m=0}^{\infty}(-1)^m\left(\frac{1}{k} + \frac{2}{\overline{d_n}} + \frac{\alpha_3}{k\overline{d_n}}\right)^m + \mathcal{O}\left(1/\overline{d_n}^2\right). \tag{4.74}$$

Now, most of the term in $\left(\frac{1}{k}+\frac{2}{\overline{d_n}}+\frac{\alpha_3}{k\overline{d_n}}\right)^m$ are of order $\mathcal{O}\left(1/\overline{d_n}^2\right)$, and we need to carefully extract the few significant terms. In the expansion of $\left(\frac{1}{k}+\frac{2}{\overline{d_n}}+\frac{\alpha_3}{k\overline{d_n}}\right)^m$ the only terms that do not contain a $1/\overline{d_n}^2$ are obtained by choosing systematically $1/k$ in each of the m factors, or choosing $1/k$ in all but one factors and either $\frac{2}{\overline{d_n}}$ or $\frac{\alpha_3}{k\overline{d_n}}$ in the last factor. There are m ways to construct these last two terms. In summary, we have for $m \geq 1$,

$$\left(\frac{1}{k}+\frac{2}{\overline{d_n}}+\frac{\alpha_3}{k\overline{d_n}}\right)^m = \frac{1}{k^m} + m\frac{\alpha_3}{k^{m-1}\overline{d_n}} + m\frac{2}{k^{m-1}\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right). \tag{4.75}$$

We can substitute (4.75)into (4.74) to get

$$\begin{aligned}\left(1+1/k+2/\overline{d_n}+\alpha_3/(k\overline{d_n})+\mathcal{O}\left(1/\overline{d_n}^2\right)\right)^{-1} &= \sum_{m=0}^{\infty}(-1)^m\frac{1}{k^m} + \sum_{m=1}^{\infty} m\frac{\alpha_3}{k^{m-1}\overline{d_n}} + m\frac{2}{k^{m-1}\overline{d_n}} + \mathcal{O}\left(1/\overline{d_n}^2\right) \\ &= \frac{1}{1+1/k} - \frac{\alpha_3}{\overline{d_n}}\frac{k}{(k+1)^2} - 2\frac{k^2}{(k+1)^2} + \mathcal{O}\left(1/\overline{d_n}^2\right) \\ &= \frac{k}{k+1} - \frac{(\alpha_3+2k)k}{\overline{d_n}(k+1)^2} + \mathcal{O}\left(1/\overline{d_n}^2\right).\end{aligned} \tag{4.76}$$

We can insert (4.76) into (4.68) to get

$$\begin{aligned}\Delta\widehat{R} &= \left(\frac{1}{k^2}+\frac{\alpha_1+\alpha_2}{\overline{d_n}k^2}+\mathcal{O}\left(1/\overline{d_n}^2\right)\right)\left(\frac{k}{k+1}-\frac{(\alpha_3+2k)k}{\overline{d_n}(k+1)^2}+\mathcal{O}\left(1/\overline{d_n}^2\right)\right) \\ &= \frac{1}{k(k+1)}+\frac{\alpha_1+\alpha_2}{\overline{d_n}k(k+1)}-\frac{(\alpha_3+2k)}{\overline{d_n}k(k+1)^2}+\mathcal{O}\left(1/\overline{d_n}^2\right),\end{aligned} \tag{4.77}$$

which leads to

$$\begin{aligned}\widehat{R}'_{uv} &= \widehat{R}_{uv} - \Delta\widehat{R} \\ &= \frac{1}{k}+\frac{2}{\overline{d_n}}+\frac{\alpha_0}{\overline{d_n}k}-\frac{1}{k(k+1)}-\frac{\alpha_1+\alpha_2}{\overline{d_n}k(k+1)}+\frac{(\alpha_3+2k)}{\overline{d_n}k(k+1)^2}+\mathcal{O}\left(1/\overline{d_n}^2\right) \\ &= \frac{1}{k+1}+\frac{2}{\overline{d_n}}+\left\{\alpha_0\frac{k+1}{k}-\frac{\alpha_1+\alpha_2}{k}+\frac{\alpha_3+2k}{k(k+1)}\right\}\frac{1}{\overline{d_n}(k+1)}+\mathcal{O}\left(1/\overline{d_n}^2\right).\end{aligned} \tag{4.78}$$

Let

$$\alpha'_0 \stackrel{\text{def}}{=} \alpha_0 + \frac{\alpha_0-\alpha_1-\alpha_2}{k} + \frac{2}{k+1} + \frac{\alpha_3}{k(k+1)}, \tag{4.79}$$

then we have

$$\widehat{R}'_{uv} = \frac{1}{k+1}+\frac{2}{\overline{d_n}}+\frac{\alpha'_0}{\overline{d_n}(k+1)}+\mathcal{O}\left(1/\overline{d_n}^2\right), \tag{4.80}$$

which matches the expression given in (4.46) for $k+1$. This completes the induction and the proof of (4.46). We note that $\alpha_0' = \alpha(k+1, u, v)$ given by (4.79) is a deterministic function of k, since $\alpha_0, \alpha_2, \alpha_2$ and α_3 all are deterministic functions of k, by the inductive hypothesis.

The cases where $u = u_{k+1}$ or $v = v_{k+1}$ are treated similarly; since they present no new difficulties they are omitted. This completes the proof of the theorem. □

Remark. *As expected* (4.79) *agrees with the update formula in the case* $k = 1$*, given by* (4.72)*.*

The update formula (4.79) implies that the **random variables** $\alpha(k_n, u, v)$ and $\alpha(k_n, u, v)/k_n$, both have bounded variation.

Corollary 22. *Let* $G_{n+1} \sim \mathcal{G}(n+1, p_n, q_n)$ *be a stochastic blockmodel with* $p_n = \omega(\log n/n)$*. Assume that* $q_n = o(p_n/n)$*. Let* G_n *be the subgraph of* G_{n+1} *induced by the vertex set* $[n]$*. Let* k_n *and* k_{n+1} *be the number of cross-community edges in* G_n *and* G_{n+1}*, respectively.*

Let u *and* v *be two vertices, and let* $\widehat{R}_{uv}^{(n)}$ *and* $\widehat{R}_{uv}^{(n+1)}$ *be the effective resistances measured in* G_n *and* G_{n+1} *respectively.*

Let $\alpha(k_n, u, v)$ *and* $\alpha(k_{n+1}, u, v)$ *be the coefficients in the expansion of* $\widehat{R}_{uv}^{(n)}$ *and* $\widehat{R}_{uv}^{(n+1)}$ *in* (4.46) *respectively. We have,*

$$|\alpha(k_{n+1}, u, v) - \alpha(k_n, u, v)| = \mathcal{O}(1) \quad \textit{with high probability,} \tag{4.81}$$

$$\tag{4.82}$$

$$\left| \frac{\alpha(k_{n+1}, u, v)}{k_{n+1}} - \frac{\alpha(k_n, u, v)}{k_n} \right| = \mathcal{O}\left(\frac{1}{k_n}\right) \quad \textit{with high probability.} \tag{4.83}$$

We start by proving that, conditioned on $k_n = k$, the **functions** $\alpha(k, u, v)$ and $\alpha(k, u, v)/k$, defined for $k \geq 1$, have bounded variation.

Proposition 1. *Let* $G_n \sim \mathcal{G}(n, p_n, q_n)$ *be a balanced, two community stochastic blockmodel with* $p_n = \omega(\log n/n)$*,* $q_n = \omega(1/n^2)$*, and* $q_n = o(p_n/n)$*. We assume that* G_n *is connected. Let* u *and*

v be two vertices. Given $k_n = k$, let $\alpha(k, u, v)$ be the coefficient in the expansion of $\widehat{R}_{uv}$ in (4.46). *Similarly, let $\alpha(k+1, u, v)$ be the corresponding quantity when $k_n = k+1$. We have*

$$|\alpha(k+1, u, v) - \alpha(k, u, v)| \leq 8, \tag{4.84}$$

$$\tag{4.85}$$

$$\left| \frac{\alpha(k+1, u, v)}{k+1} - \frac{\alpha(k, u, v)}{k} \right| \leq \frac{6}{k}. \tag{4.86}$$

Proof. The proof is a direct consequence of (4.79), and the fact that $|\alpha(k, u, v)| \leq 2k$.

Let us start with the first inequality. From (4.79) we have

$$\alpha(k+1, u, v) - \alpha(k, u, v) = \frac{\alpha(k, u, v) - \alpha_1 - \alpha_2}{k} + \frac{2}{k+1} + \frac{\alpha_3}{k(k+1)}, \tag{4.87}$$

and thus

$$|\alpha(k+1, u, v) - \alpha(k, u, v)| \leq \frac{6k}{k} + \frac{2}{k+1} + \frac{2k}{k(k+1)} = 6 + \frac{4}{k+1} \leq 8. \tag{4.88}$$

We now show the second inequality. Again, from (4.79) we have

$$\begin{aligned} \frac{\alpha(k+1, u, v)}{k+1} - \frac{\alpha(k, u, v)}{k} &= \frac{\alpha(k, u, v)}{k+1} - \frac{\alpha(k, u, v)}{k} + \frac{\alpha(k, u, v) - \alpha_1 - \alpha_2}{k(k+1)} + \frac{2}{(k+1)^2} + \frac{\alpha_3}{k(k+1)^2} \\ &= -\frac{\alpha(k, u, v)}{k(k+1)} + \frac{\alpha(k, u, v) - \alpha_1 - \alpha_2}{k(k+1)} + \frac{2}{(k+1)^2} + \frac{\alpha_3}{k(k+1)^2} \\ &= -\frac{\alpha_1 + \alpha_2}{k(k+1)} + \frac{2}{(k+1)^2} + \frac{\alpha_3}{k(k+1)^2}, \end{aligned} \tag{4.89}$$

and thus

$$\begin{aligned} \left| \frac{\alpha(k+1, u, v)}{k+1} - \frac{\alpha(k, u, v)}{k} \right| &\leq \frac{4k}{k(k+1)} + \frac{2}{(k+1)^2} + \frac{2k}{k(k+1)^2} = \frac{4}{k+1} + \frac{2}{(k+1)^2} + \frac{2}{(k+1)^2} \\ &\leq \frac{4}{k+1} \left\{ 1 + \frac{1}{k+1} \right\} \leq \frac{6}{k+1} \leq \frac{6}{k}. \end{aligned} \tag{4.90}$$

□

Proof of corollary 22. We first verify that with high probability $k_{n+1} - k_n$ is bounded. We then apply proposition 1.

Let $\Delta k_n \stackrel{\text{def}}{=} k_{n+1} - k_n$ be the number of adjacent cross-community edges that have vertex $n+1$ as one of their endpoint. Δk_n is a binomial random variable $B(n_1, q_n)$, where we assume without loss of generality that vertex $n+1 \in C_2$. Because $\mathbb{E}[\Delta k_n] = nq_n/2 = p_n o(1)$, $\mathbb{E}[\Delta k_n]$ is bounded, and there exists κ such that $\forall n, \mathbb{E}[\Delta k_n] < \kappa$. Using a Chernoff bound we have

$$\text{Prob}\left(|\mathbb{E}[\Delta k_n] - \Delta k_n| > 3\sqrt{\kappa}\right) \leq 2\exp\left(-\frac{1}{3}\right) < 0.01, \tag{4.91}$$

and thus

$$|k_{n+1} - k_n| = |k_{n+1} - \mathbb{E}[\Delta k_n] + \mathbb{E}[\Delta k_n] - k_n| < 6\sqrt{\kappa} \quad \text{with probability ¿ 0.99} \tag{4.92}$$

In other words, with high probability $k_{n+1} - k_n$ is bounded by $C = \lceil 6\sqrt{\kappa} \rceil$, independently of n.

Finally, we have

$$\begin{aligned} \left|\frac{\alpha(k_{n+1},u,v)}{k_{n+1}} - \frac{\alpha(k_n,u,v)}{k_n}\right| &= \left|\sum_{k=k_n}^{k_{n+1}-1} \frac{\alpha(k+1,u,v)}{k+1} - \frac{\alpha(k,u,v)}{k}\right| \\ &\leq \sum_{k=k_n}^{k_{n+1}-1} \left|\frac{\alpha(k+1,u,v)}{k+1} - \frac{\alpha(k,u,v)}{k}\right| \end{aligned} \tag{4.93}$$

Because of corollary 1, each term in the sum is bounded by $6/k$; the largest upper bound being $6/k_n$. Also, with high probability there are at most C terms, and thus

$$\left|\frac{\alpha(k_{n+1},u,v)}{k_{n+1}} - \frac{\alpha(k_n,u,v)}{k_n}\right| \leq C\frac{6}{k_n} \tag{4.94}$$

□

We now combine lemma 15 with theorem 4 to estimate the perturbation created by the addition of an $n+1^{\text{th}}$ vertex to G_n. The following lemma shows that adding an additional vertex, with corresponding edges to either one of the communities does not change the effective resistance, as long as no new cross-community edges are created.

Lemma 23. *Let $G_{n+1} \sim \mathcal{G}(n+1, p_n, q_n)$ be a stochastic blockmodel with $p_n = \omega(\log n/n)$. Assume that $q_n = o(p_n/n)$. Let G_n be the subgraph of G_{n+1} induced by the vertex set $[n]$. Let k_n and k_{n+1} be the number of cross-community edges in G_n and G_{n+1}, respectively.*

Let u and v be two vertices in G_{n+1}, for which the effective resistance $\widehat{R}_{uv}^{(n)}$, measured in G_n, is properly defined. The corresponding effective resistance measured in G_{n+1} is then given by

$$\widehat{R}_{u,v}^{(n)} - \widehat{R}_{uv}^{(n+1)} = \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right) \geq 0, \quad \textit{if } u \textit{ and } v \textit{ belong to the same community.} \tag{4.95}$$

If u and v are in different communities, we have the following inequalities,

$$\begin{cases} \widehat{R}_{u,v}^{(n)} - \widehat{R}_{uv}^{(n+1)} = \mathcal{O}\left(\dfrac{1}{\overline{d_n}^2}\right) \geq 0, & \textit{if} \quad k_n = k_{n+1}, \\ \\ \widehat{R}_{u,v}^{(n)} - \widehat{R}_{uv}^{(n+1)} \geq \dfrac{2}{k_n^2} + \dfrac{1}{\overline{d_n}}\mathcal{O}\left(\dfrac{1}{k_n}\right) \geq 0, & \textit{if} \quad k_{n+1} > k_n. \end{cases} \tag{4.96}$$

Proof. Because of lemma 15, we have

$$\frac{1}{\overline{d_{n+1}}} = \frac{1}{\overline{d_n}} + \mathcal{O}\left(\frac{1}{\overline{d_{n+1}}^2}\right), \tag{4.97}$$

and

$$\frac{1}{\overline{d_{n+1}}^2} = \frac{1}{\overline{d_n}^2} + \mathcal{O}\left(\frac{1}{\overline{d_{n+1}}^3}\right). \tag{4.98}$$

If u and v are in the same community, the expression for $\widehat{R}_{u,v}^{(n+1)}$ and $\widehat{R}_{u,v}^{(n)}$, given by (4.46) coincide, up to order $\mathcal{O}\left(1/\overline{d_n}^2\right)$.

When u and v are in different communities, we need to consider the values of k_n and k_{n+1}. If $k_n = k_{n+1}$ then $\alpha(k_n, u, v) = \alpha(k_{n+1}, u, v)$, and thus $\widehat{R}_{u,v}^{(n+1)} = \widehat{R}_{u,v}^{(n)}$, up to order $\mathcal{O}\left(1/\overline{d_n}^2\right)$.

If $k_{n+1} > k_n$, we will show that the decrease in effective resistance is of order $\Theta\left(1/k_{n+1}^2\right)$. We first recall that $\widehat{R}_{u,v}^{(n)} \geq \widehat{R}_{uv}^{(n+1)}$, since G^{n+1} has more edges than G_n, and thus

$$|\widehat{R}_{u,v}^{(n)} - \widehat{R}_{uv}^{(n+1)}| = \widehat{R}_{u,v}^{(n)} - \widehat{R}_{uv}^{(n+1)}. \tag{4.99}$$

Using the expression for $\widehat{R}_{u,v}^{(n+1)}$ and $\widehat{R}_{u,v}^{(n)}$, given by (4.46) we have

$$\begin{aligned}
\left|\widehat{R}_{u,v}^{(n)} - \widehat{R}_{uv}^{(n+1)}\right| &= \widehat{R}_{u,v}^{(n)} - \widehat{R}_{uv}^{(n+1)} \\
&= \frac{1}{k_n} - \frac{1}{k_{n+1}} + \frac{\alpha(k_n,u,v)}{\overline{d_n}k_n} - \frac{\alpha(k_{n+1},u,v)}{\overline{d_{n+1}}k_{n+1}} + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right) \\
&\geq \frac{1}{k_n} - \frac{1}{k_n+1} + \frac{\alpha(k_n,u,v)}{\overline{d_n}k_n} - \frac{\alpha(k_{n+1},u,v)}{\overline{d_n}(k_n+1)} + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right) \\
&\geq \frac{1}{k_n(k_n+1)} + \frac{1}{\overline{d_n}}\left(\frac{\alpha(k_n,u,v)}{k_n} - \frac{\alpha(k_{n+1},u,v)}{k_n+1}\right) + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right). \\
&\geq \frac{1}{2k_n^2} + \frac{1}{\overline{d_n}}\left(\frac{\alpha(k_n,u,v)}{k_n} - \frac{\alpha(k_{n+1},u,v)}{k_n+1}\right) + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right).
\end{aligned} \tag{4.100}$$

We recall that corollary 22 implies that

$$\left|\frac{\alpha(k_n+1,u,v)}{k_n+1} - \frac{\alpha(k_n,u,v)}{k_n}\right| = \mathcal{O}\left(\frac{1}{k_n^2}\right) \tag{4.101}$$

and thus

$$\left|\widehat{R}_{u,v}^{(n)} - \widehat{R}_{uv}^{(n+1)}\right| \geq \frac{1}{2k_n^2} + \frac{1}{\overline{d_n}}\mathcal{O}\left(\frac{1}{k_n}\right), \tag{4.102}$$

which completes the proof. □

4.8.3.3 The Distance D_n Under the Null Hypothesis

Theorem 5. *Let $G_{n+1} \sim \mathcal{G}(n+1, p_n, q_n)$ be a stochastic blockmodel with $p_n = \omega(\log n/n)$, $q_n = \omega\left(1/n^2\right)$, and $q_n = o(p_n/n)$. Let G_n be the subgraph induced by the vertex set $[n]$, and let $D_n = \mathrm{RD}(G_n, G_{n+1})$ be the normalized effective resistance distance,* RD*, defined in* (4.3).

Suppose that the introduction of $n+1$ does not create additional cross-community edges, that is $k_n = k_{n+1}$, then

$$D_n - h(n,k_n) = \mathcal{O}\left(\frac{n^2}{\overline{d_n}^2}\right) \geq 0, \tag{4.103}$$

where

$$h(n,k_n) = \left\lfloor \frac{n}{2} \right\rfloor + \left\lceil \frac{n}{2} \right\rceil \frac{k_n}{1+k_n}. \tag{4.104}$$

Proof. Since vertex $n+1$ is isolated in G_n, the change in resistance at vertex $n+1$ between G_n and G_{n+1} will behave quite differently than as described in Theorem 3. For this reason, we separate the renormalized resistance distance into two portions: the pair of nodes that do and do not contain vertex $n+1$,

$$D_n = \sum_{u<v\leq n+1} \left| R_{u,v}^{(n+1)} - R_{u,v}^{(n)} \right| = \sum_{u<v\leq n} \left| R_{u,v}^{(n+1)} - R_{u,v}^{(n)} \right| + \sum_{u\leq n} \left| R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)} \right|. \tag{4.105}$$

Let us first study the second sum. Because vertex $n+1$ is isolated at time n, $R_{u,n+1}^{(n)} = 1$. If u and $n+1$ are in the same community, then $R_{u,n+1}^{(n+1)} \leq \widehat{R}_{uv}^{(n)} = \mathcal{O}\left(1/\overline{d_n}^2\right)$, and is therefore negligible. We can use the trivial bound $R_{u,n+1}^{(n+1)} \geq 0$ that yields

$$\left| R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)} \right| \leq 1.$$

If u and $n+1$ are in different communities, then a tighter bound can be derived by considering the bottleneck formed by the cross-community edges. Indeed, a coarse application of Nash-Williams – using only the cross-community cut-set – tells us that the effective resistance between vertices in different communities is bounded by $1/k_n$, and so the renormalized resistance is bounded by

$$R_{u,n+1}^{(n+1)} \geq \frac{1}{1+k_n},$$

which implies

$$\left| R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)} \right| \leq \frac{k_n}{1+k_n}.$$

Observing that there are $\left\lfloor \frac{n}{2} \right\rfloor$ possible in-community connections and $\left\lceil \frac{n}{2} \right\rceil$ possible cross-community connections, we have

$$\sum_{u\leq n} \left| R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)} \right| \leq \left\lfloor \frac{n}{2} \right\rfloor + \left\lceil \frac{n}{2} \right\rceil \frac{k_n}{1+k_n}. \tag{4.106}$$

We now consider the first sum in (4.105). Corollary 23 combined with lemma 16 yield

$$\left| R_{u,v}^{(n+1)} - R_{u,v}^{(n)} \right| \leq \left| \widehat{R}_{u,v}^{(n+1)} - \widehat{R}_{u,v}^{(n)} \right| = \mathcal{O}\left(\frac{1}{\overline{d_n}^2} \right), \tag{4.107}$$

which leads to the following bound on the first sum,

$$\sum_{u<v\leq n} \left| R_{u,v}^{(n+1)} - R_{u,v}^{(n)} \right| = \mathcal{O}\left(\frac{n^2}{\overline{d_n}^2} \right). \tag{4.108}$$

Combining (4.106) and (4.108) yields

$$0 \leq D_n = \sum_{u<v\leq n+1} \left| R_{u,v}^{(n+1)} - R_{u,v}^{(n)} \right| \leq \left\lfloor \frac{n}{2} \right\rfloor + \left\lceil \frac{n}{2} \right\rceil \frac{k_n}{1+k_n} + \mathcal{O}\left(\frac{n^2}{\overline{d_n}^2} \right) = h(n,k_n) + \mathcal{O}\left(\frac{n^2}{\overline{d_n}^2} \right), \quad (4.109)$$

which implies the advertised result. □

The following corollary, which is an immediate consequence of the previous theorem provides the appropriate renormalization of $D_n - h(n,k_n)$ under the null hypothesis H_0.

Corollary 24. *Let $G_{n+1} \sim \mathcal{G}(n+1, p_n, q_n)$ be a stochastic blockmodel with the same conditions on p_n and q_n as in Theorem 5, and let G_n be the subgraph induced by the vertex set $[n]$.*

Suppose that the introduction of $n+1$ does not create additional cross-community edges, that is $k_n = k_{n+1}$, then

$$0 \leq p_n^2 \left(D_n - h(n,k_n) \right) = \mathcal{O}(1) \quad (4.110)$$

Proof. As explained in lemma 15, we assume without loss of generality that n is even. We have then

$$\overline{d_n}^2 = \frac{(n/2-1)^2}{p_n^2} \quad (4.111)$$

and thus

$$\frac{n^2}{\overline{d_n}^2} = \frac{n^2}{p_n^2 (n/2-1)^2} = \frac{4}{p_n^2}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right), \quad (4.112)$$

which leads to

$$\frac{n^2 p_n^2}{\overline{d_n}^2} = 4\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right). \quad (4.113)$$

We recall that theorem 5 gives us the following bound on $(D_n - h(n,k_n))$ under the null hypothesis,

$$(D_n - h(n,k_n)) = \mathcal{O}\left(\frac{n^2}{\overline{d_n}^2} \right), \quad (4.114)$$

we conclude that

$$p_n^2 \left(D_n - h(n,k_n) \right) = \mathcal{O}(1). \quad (4.115)$$

□

4.8.3.4 The Distance D_n Under the Alternate Hypothesis

We now consider the case where the addition of node $n+1$ leads to an increase in the number of cross-community edges. Loosening the bottleneck between the two communities creates a significant change in the normalized effective resistance distance between G_n and G_{n+1}.

Theorem 6. *Let $G_{n+1} \sim \mathcal{G}(n+1, p_n, q_n)$ be a stochastic blockmodel with $p_n = \omega(\log n/n)$, $q_n = \omega\left(1/n^2\right)$, and $q_n = o(p_n/n)$. Let G_n be the subgraph induced by the vertex set $[n]$.*

Suppose that the introduction of $n+1$ creates additional cross-community edges, that is $k_{n+1} > k_n$, then

$$0 \leq \frac{1}{16}\left\{\frac{n^2}{k_n^2} + \frac{n^2}{\overline{d_n}}\mathcal{O}\left(\frac{1}{k_n}\right)\right\} \leq D_n - h(n,k_n) \leq \frac{n^2}{k_n^2} + \mathcal{O}\left(\frac{n^2}{\overline{d_n}^2}\right), \tag{4.116}$$

where $h(n,k_n)$ is defined in (4.104).

Proof. As before, we split the distance D_n into two terms,

$$D_n = \sum_{u<v\leq n+1}\left|R_{u,v}^{(n+1)} - R_{u,v}^{(n)}\right| = \sum_{u<v\leq n}\left|R_{u,v}^{(n+1)} - R_{u,v}^{(n)}\right| + \sum_{1\leq i\leq n}\left|R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)}\right|. \tag{4.117}$$

Again, we analyze the second sum, which will generate the same linear contribution,

$$\left|R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)}\right| = 1 - R_{u,n+1}^{(n+1)} = \frac{1}{1+\widehat{R}_{u,n+1}^{(n+1)}}.$$

Because we seek an **upper** bound on $R_{u,n+1}^{(n+1)}$ to obtain an **lower** bound on the change $\left|R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)}\right|$ we have a more refined analysis of $R_{u,n+1}^{(n+1)}$.

In the case where u and $n+1$ are in the same community, Theorem 4 tells us that

$$\widehat{R}_{u,n+1}^{(n+1)} = \frac{2}{\overline{d_{n+1}}} + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right),$$

We use the inequality $\frac{1}{1+x} \geq 1-x$, which is valid for all $x > -1$, to get a lower bound,

$$\left|R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)}\right| = \frac{1}{1+\widehat{R}_{u,n+1}^{(n+1)}} \geq 1 - \frac{2}{\overline{d_{n+1}}} + \mathcal{O}\left(\frac{1}{\overline{d_n}^2}\right) = 1 + \mathcal{O}\left(\frac{1}{\overline{d_n}}\right). \tag{4.118}$$

We also use the inequality $\frac{1}{1+x} \leq 1 - x/2$, which is valid for all $x \in [0,1]$, to get an upper bound,

$$\left|R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)}\right| = \frac{1}{1+\widehat{R}_{u,n+1}^{(n+1)}} \leq 1 - \frac{1}{\overline{\overline{d_{n+1}}}} + \mathcal{O}\left(\frac{1}{\overline{\overline{d_n}}^2}\right) = 1 + \mathcal{O}\left(\frac{1}{\overline{\overline{d_n}}}\right). \tag{4.119}$$

If u and $n+1$ are in separate communities, Lemma 20 tells us that

$$\frac{1}{k_{n+1}} \leq \widehat{R}_{u,n+1}^{(n+1)} \leq \frac{1}{k_{n+1}} + \frac{4}{\overline{\overline{d_{n+1}}}} + \mathcal{O}\left(\frac{1}{\overline{\overline{d_{n+1}}}^2}\right) \leq \frac{1}{1+k_n} + \frac{4}{\overline{\overline{d_{n+1}}}} + \mathcal{O}\left(\frac{1}{\overline{\overline{d_n}}^2}\right). \tag{4.120}$$

Using again the inequality $\frac{1}{1+x} \geq 1 - x$, we get a lower bound,

$$\left|R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)}\right| \geq 1 - \frac{1}{1+k_n} - \frac{4}{\overline{\overline{d_{n+1}}}} + \mathcal{O}\left(\frac{1}{\overline{\overline{d_n}}^2}\right) = \frac{k_n}{1+k_n} + \mathcal{O}\left(\frac{1}{\overline{\overline{d_n}}}\right), \tag{4.121}$$

and using the inequality $\frac{1}{1+x} \leq 1 - x/2$ we get an upper bound,

$$\left|R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)}\right| \leq 1 - \frac{1}{2k_{n+1}} \leq 1. \tag{4.122}$$

Combining (4.118), (4.119),(4.121), and (4.122), we get

$$\left\lfloor \frac{n}{2} \right\rfloor + \left\lceil \frac{n}{2} \right\rceil \frac{k_n}{1+k_n} + \mathcal{O}\left(\frac{n}{\overline{\overline{d_n}}}\right) \leq \sum_{u \leq n} \left|R_{u,n+1}^{(n+1)} - R_{u,n+1}^{(n)}\right| \leq n + \mathcal{O}\left(\frac{n}{\overline{\overline{d_n}}}\right). \tag{4.123}$$

We now consider the first sum in (4.117). To get lower and upper bounds on $\left|\widehat{R}_{u,v}^{(n+1)} - \widehat{R}_{u,v}^{(n)}\right|$ we use lemma 16.

We first observe that for n sufficiently large, we have $\widehat{R}_{u,v}^{(n)} \leq 1$, and thus $\widehat{R}_{u,v}^{(n+1)} \leq 1$. Combining this upper bound on the effective resistance with lemma 16 we get

$$\text{if} \quad C(u,v) \leq \left|\widehat{R}_{u,v}^{(n+1)} - \widehat{R}_{u,v}^{(n)}\right| \quad \text{then} \quad \frac{C(u,v)}{4} \leq \frac{C(u,v)}{(1+\widehat{R}_{u,v}^{(n)})(1+\widehat{R}_{u,v}^{(n+1)})} \leq \left|\widehat{R}_{u,v}^{(n+1)} - \widehat{R}_{u,v}^{(n)}\right|. \tag{4.124}$$

From corollary 23 we have

$$\begin{cases} C(u,v) = \mathcal{O}\left(\frac{1}{\overline{\overline{d_n}}^2}\right) & \text{if } u \text{ and } v \text{ are in the same community,} \\ C(u,v) \geq \frac{1}{k_n^2} + \frac{1}{\overline{\overline{d_n}}}\mathcal{O}\left(\frac{1}{k_n}\right) & \text{otherwise,} \end{cases} \tag{4.125}$$

and therefore

$$4\sum_{u<v\leq n} \left|R_{u,v}^{(n+1)} - R_{u,v}^{(n)}\right| \geq \sum_{\substack{u,v \in \text{different} \\ \text{communities}}} C(u,v) \; + \sum_{\substack{u,v \in \text{same} \\ \text{community}}} C(u,v) \geq \frac{n^2}{4k_n^2} + \frac{n^2}{4\overline{\overline{d_n}}}\mathcal{O}\left(\frac{1}{k_n}\right) + \frac{n^2}{4}\mathcal{O}\left(\frac{1}{\overline{\overline{d_n}}^2}\right), \tag{4.126}$$

where the differences between $n/2$ and the exact size of C_1 or C_2 are absorbed in the error terms. Also, we have

$$\frac{1}{\overline{d_n}^2} = \frac{1}{\overline{d_n}}\mathcal{O}\left(\frac{1}{k_n}\right), \tag{4.127}$$

and thus

$$\sum_{u<v\leq n}\left|R_{u,v}^{(n+1)} - R_{u,v}^{(n)}\right| \geq \frac{1}{16}\left\{\frac{n^2}{k_n^2} + \frac{n^2}{\overline{d_n}}\mathcal{O}\left(\frac{1}{k_n}\right)\right\}. \tag{4.128}$$

Finally, we note that

$$\frac{n}{\overline{d_n}} = \frac{n^2}{\overline{d_n}k_n}\frac{k_n}{n} = \frac{n^2}{\overline{d_n}k_n}\frac{k_n}{n} \tag{4.129}$$

Because of lemma 19, we have asymptotically with high probability,

$$\frac{n}{\overline{d_n}} = \frac{n^2}{\overline{d_n}k_n}\mathcal{O}\left(\frac{\mathbb{E}\left[k_n\right]}{n}\right) = \frac{n^2}{\overline{d_n}k_n}\mathcal{O}\left(nq\right) = \frac{n^2}{\overline{d_n}}\mathcal{O}\left(\frac{1}{k_n}\right), \tag{4.130}$$

and thus we conclude that

$$\mathcal{O}\left(\frac{n}{\overline{d_n}}\right) = \frac{n^2}{\overline{d_n}}\mathcal{O}\left(\frac{1}{k_n}\right). \tag{4.131}$$

Lastly, we add the two sums (4.123) and (4.128) to get

$$\left\lfloor\frac{n}{2}\right\rfloor + \left\lceil\frac{n}{2}\right\rceil\frac{k_n}{1+k_n} + \frac{1}{16}\left\{\frac{n^2}{k_n^2} + \frac{n^2}{\overline{d_n}}\mathcal{O}\left(\frac{1}{k_n}\right)\right\} \leq \sum_{u<v\leq n+1}\left|R_{u,v}^{(n+1)} - R_{u,v}^{(n)}\right|. \tag{4.132}$$

The leading term linear term, $h(n,k_n)$, in (4.132) can be subtracted to arrive at the advertised result. □

Using the same normalization described in corollary 24 we obtain a very different growth for $p_n^2(D_n - h(n,k_n))$ in the case of the alternate hypothesis.

Corollary 25. *Let $G_{n+1} \sim \mathcal{G}(n+1, p_n, q_n)$ be a stochastic blockmodel with the same conditions on p_n and q_n as in Theorem 5, and let G_n be the subgraph induced by the vertex set $[n]$.*

Suppose that the introduction of $n+1$ creates additional cross-community edges, that is $k_{n+1} > k_n$, then

$$0 \leq p_n\left(D_n - h(n,k_n)\right) \to \infty \quad \textit{with high probability.} \tag{4.133}$$

Proof. As explained in lemma 15, we assume without loss of generality that n is even. From (4.116) we have

$$p_n \left(D_n - h(n, k_n)\right) \geq \frac{1}{16} \left\{ \frac{(np_n)^2}{k_n^2} + \frac{(np_n)^2}{\overline{d_n}} \mathcal{O}\left(\frac{1}{k_n}\right) \right\} \tag{4.134}$$

Without loss of generality we assume n even, and we have

$$\frac{(np_n)^2}{\overline{d_n}} \mathcal{O}\left(\frac{1}{k_n}\right) = \frac{(np_n)^2}{p_n(n/2-1)q_n(n^2/4)} \mathcal{O}\left(1\right) = \frac{p_n}{nq_n} \mathcal{O}\left(1\right) = \omega\left(1\right) \mathcal{O}\left(1\right). \tag{4.135}$$

Therefore the second term cannot go to $-\infty$. We will prove that the first term goes to infinity. We have

$$\frac{np_n}{k_n} = \frac{np_n}{\mathbb{E}\left[k_n\right]} \frac{\mathbb{E}\left[k_n\right]}{k_n} = \frac{np_n}{q_n(n/2)^2} \frac{\mathbb{E}\left[k_n\right]}{k_n} = \frac{4p_n}{nq_n} \frac{\mathbb{E}\left[k_n\right]}{k_n}. \tag{4.136}$$

From lemma 19 we know that asymptotically $\mathbb{E}\left[k_n\right]/k_n = \Theta\left(1\right)$ with high probability. Also, we have $p_n/(nq_n) = \omega\left(1\right)$. This concludes the proof. □

The quantity $p_n^2(D_n - h(n, k_n))$ could provide a statistic to test the null hypothesis $k_n = k_{n+1}$ against the alternate hypothesis $k_n < k_{n+1}$. Unfortunately, computing $p_n^2(D_n - h(n, k_n))$ requires the knowledge of the unknown parameter p_n, and unknown variable $h(n, k_n)$. We therefore propose two estimates that converge to these unknowns. A simple estimate of $h(n, k_n)$ is provided by n. Since we assume that there are much fewer cross-community edges than edges within each community, we can estimate p_n from the total number of edges.

We start with two technical lemmas. The first lemma shows that can replace $h(n, k_n)$ with n.

Lemma 26. *Let $G_n \sim \mathcal{G}(n, p_n, q_n)$ be a stochastic blockmodel with the same conditions on p_n and q_n as in Theorem 5. If $p_n = \mathcal{O}\left(1/\sqrt{n}\right)$, then we have*

$$\lim_{n\to\infty} p_n^2 \left(n - h(n, k_n)\right) = 0 \quad \textit{with high probability.} \tag{4.137}$$

Proof. We have

$$n - h(n, k_n) = \left\lceil \frac{n}{2} \right\rceil \frac{1}{k_n + 1}. \tag{4.138}$$

Because $\lceil n/2 \rceil = (n/2)\Theta(1)$, we have

$$p_n^2(n - h(n, k_n)) = \Theta(1) \frac{np_n^2}{2(k_n + 1)} = \frac{np_n^2}{2\mathbb{E}[k_n]} \frac{\mathbb{E}[k_n]}{k_n} \frac{k_n}{k_n + 1} \Theta(1). \tag{4.139}$$

Now, we have $k_n/(k_n + 1) < 1$, $np_n^2 = \mathcal{O}(1)$, and $\mathbb{E}[k_n] = \omega(1)$, therefore

$$\lim_{n\to\infty} \frac{np_n^2}{2\mathbb{E}[k_n]} \frac{k_n}{k_n + 1} \Theta(1) = 0. \tag{4.140}$$

Finally, we recall that $\mathbb{E}[k_n]/k_n = \Theta(1)$ with high probability, which concludes the proof. □

We now consider the estimation of p_n.

Lemma 27. *Let $G_n \sim \mathcal{G}(n, p_n, q_n)$ be a stochastic blockmodel with the same conditions on p_n and q_n as in Theorem 5. Let m_n be the total number of edges in G_n. Then the probability p_n can be estimated asymptotically from m_n and n,*

$$\frac{4m_n}{n^2} = p_n(1 + \mathcal{O}(1/n)), \quad \textit{with high probability.} \tag{4.141}$$

Proof. The proof proceeds in two steps. We first show that k_n concentrates around its expectation $\mathbb{E}[k_n]$, and then we argue that $\lim_{n\to\infty} 4\mathbb{E}[m_n]/n^2 = p_n$.

The total number of edges, m_n, in the graph G_n, can be decomposed as

$$m_n = m_{n_1} + m_{n_2} + k_n, \tag{4.142}$$

where m_{n_1} (m_{n_2}) is the number of edges in community C_1 (C_2). The three random variables are binomial (with different parameters), and they concentrate around their respective expectations. Because the sum of these binomials is finite, m_n also concentrates around its expectation. Consequently, we can combine three Chernoff inequalities using a union bound, and show that

$$\frac{m_n}{\mathbb{E}[m_n]} = \Theta(1), \quad \text{with high probability.} \tag{4.143}$$

A quick computation of $\mathbb{E}[m_n]$ shows that

$$\mathbb{E}[m_n] = p_n \frac{n^2}{4} \left(1 - \frac{2}{n} + \frac{q_n}{p_n} + o\left(\frac{1}{n^2}\right)\right). \tag{4.144}$$

Also, $q_n/p_n = o\left(1/n\right)$, and thus

$$\mathbb{E}\left[m_n\right] = p_n \frac{n^2}{4}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right). \tag{4.145}$$

To conclude, we combine (4.143) and (4.145), to get

$$\frac{4m_n}{n^2} = \frac{m_n}{\mathbb{E}\left[m_n\right]}\frac{4\mathbb{E}\left[m_n\right]}{n^2} = p_n\left(1 - \mathcal{O}\left(\frac{1}{n}\right)\right), \tag{4.146}$$

which concludes the proof. □

We define the following statistic that asymptotically converges toward $p_n^2(D_n - h(n.k_n))$ with high probability, as explained in the next theorem.

Definition 8. *Let $G_{n+1} \sim \mathcal{G}(n, p_n, q_n)$ be a stochastic blockmodel with the same conditions on p_n and q_n as in Theorem 5. Let G_n be the subgraph induced by the vertex set $[n]$. Let $D_n = \mathrm{RD}\left(G_n, G_{n+1}\right)$ be the normalized effective resistance distance,* RD*, defined in* (4.3)*.*

We define the statistic

$$Z_n \overset{def}{=} \frac{16m_n^2}{n^4}\left(D_n - n\right). \tag{4.147}$$

Theorem 7. *Let $G_n \sim \mathcal{G}(n, p_n, q_n)$ be a stochastic blockmodel with the same conditions on p_n and q_n as in Theorem 5. If $p_n = \mathcal{O}\left(1/\sqrt{n}\right)$, then we have*

$$Z_n = p_n^2\left(D_n - h(n, k_n)\right)\left(1 + o\left(1\right)\right), \qquad \textit{with high probability.} \tag{4.148}$$

Proof. The proof is an elementary consequence of the two lemmas 26 and 27. We have

$$\frac{16m_n^2}{n^2}\left(D_n - n\right) = \frac{16m_n^2}{n^4}\left(D_n - h(n, k_n)\right) + \frac{16m_n^2}{n^4}\left(h(n, k_n) - n\right) \tag{4.149}$$

Using lemma 27, we have

$$\frac{16m_n^2}{n^2}\left(D_n - n\right) = p_n^2\left(1 + \mathcal{O}\left(1/n\right)\right)^2\left(D_n - h(n, k_n)\right) + \left(1 + \mathcal{O}\left(1/n\right)\right)^2 p_n^2\left(h(n, k_n) - n\right) \tag{4.150}$$

Lemma 26 shows that the second term can be neglected,

$$\begin{aligned}\frac{16m_n^2}{n^2}\left(D_n - n\right) &= p_n^2\left(1+\mathcal{O}\left(1/n\right)\right)\left(D_n - h(n,k_n)\right) + \left(1+\mathcal{O}\left(1/n\right)\right)o\left(1\right) \\ &= p_n^2\left(D_n - h(n,k_n)\right)\left(1+\mathcal{O}\left(1/n\right)\right) + o\left(1\right).\end{aligned} \tag{4.151}$$

Because $p_n^2\left(D_n - h(n,k_n)\right)$ is either bounded, or goes to infinity, we have

$$\frac{(16m_n^2/n^2)(D_n - n)}{p_n^2\left(D_n - h(n,k_n)\right)} = 1 + o\left(1\right), \tag{4.152}$$

which concludes the proof. □

We finally arrive at the main theorem.

Theorem 8. *Let $G_{n+1} \sim \mathcal{G}(n, p_n, q_n)$ be a stochastic blockmodel with the same conditions on p_n and q_n as in Theorem 5. Let G_n be the subgraph induced by the vertex set $[n]$.*

To test the hypothesis

$$H_0: \quad k_n = k_{n+1} \tag{4.153}$$

versus

$$H_1: \quad k_n < k_{n+1} \tag{4.154}$$

we use the test based on the statistic Z_n defined in (4.147) where we accept H_0 if $Z_n < z_\alpha$ and accept H_1 otherwise. The threshold z_α for the rejection region satisfies

$$\mathrm{Prob}_{H_0}\left(Z_n \geq z_\alpha\right) \leq \alpha \quad \textit{as} \quad n \to \infty, \tag{4.155}$$

and

$$\mathrm{Prob}_{H_1}\left(Z_n \geq z_\alpha\right) \to 1 \quad \textit{as} \quad n \to \infty. \tag{4.156}$$

The test has therefore asymptotic level α and asymptotic power 1.

Proof. Assume H_0 to be true. Because of corollary 24 and Theorem 7,

$$Z_n = \mathcal{O}\left(1\right), \quad \text{with high probability.} \tag{4.157}$$

In other words, for every $0 < \alpha < 1$ there exists z_α such that

$$\text{Prob}\left(Z_n < z_\alpha\right) = 1 - \alpha, \quad \text{as} \quad n \to \infty, \tag{4.158}$$

or

$$\text{Prob}\left(Z_n \geq z_\alpha\right) = \alpha, \quad \text{as} \quad n \to \infty. \tag{4.159}$$

Assume that H_1 is true. Because of corollary 25 and Theorem 7,

$$Z_n = \omega\left(1\right), \quad \text{with high probability} \tag{4.160}$$

Therefore, for every $0 < \gamma < 1$, there exists n_0 such that

$$\forall n \geq n_0, \text{Prob}\left(Z_n > z_\alpha\right) = 1 - \gamma, \quad \text{as} \quad n \to \infty. \tag{4.161}$$

In other words,

$$\text{Prob}_{H_1}\left(Z_n \geq z_\alpha\right) \to 1 \quad \text{as} \quad n \to \infty, \tag{4.162}$$

which concludes the proof. □

Chapter 5

Distances for Efficient Graph Comparison: a Practitioner's Guide

This chapter written in collaboration with François Meyer. [1] [2]

5.1 Introduction

In the era of big data, comparison and matching are ubiquitous tasks. A graph is a particular type of data structure which records the interactions between some collection of agents.[3] This type of data structure relates connections between objects, rather than directly relating the properties of those objects. The interconnectedness of the object in graph data disallows many common statistical techniques used to analyze tabular datasets. The need for new analytical techniques for visualizing, comparing, and understanding graph data has given rise to a rich field of study [41].

In this work, we focus on tools for pairwise comparison of graphs. Such comparison often takes place within the contest of anomaly detection and graph matching. In the former, one has a sequence of graphs (often a time series) and hopes to establish at what time steps the graphs change "significantly" at any given timestep. In the latter, one has a collection of graphs, and wants to establish whether a sample is likely to have been drawn from that collection. Both problems require the ability to effectively compare two graphs. However, the utility of any given comparison method varies with the type of information the user is looking for; one may care primarily about large scale graph features such as community structure or the existence of highly connected "hubs";

[1] Department of Electrical and Computer Engineering, University of Colorado, Boulder, CO, 80309
[2] Department of Applied Mathematics, University of Colorado, Boulder, CO, 80309
[3] These objects are sometimes referred to as "complex networks;" we will use the mathematician's term "graph" throughout the chapter.

or, one may be focused on smaller scale structure such as local connectivity (i.e. the degree of a vertes) or the ubiquity of substructures such as triangles.

Existing surveys of graph distances are limited to observational datasets [42]. While authors try to choose datasets that are exemplars of certain classes of networks (e.g. social, biological, or computer networks), it is difficult to generalize these studies to other datasets.

In this chapter, we take a different approach. We consider existing ensembles of random graphs as prototypical examples of certain graph **structures**, which are the building blocks of existing empirical network datasets. We propose therefore to study the ability of various distances to compare two samples randomly drawn from distinct ensembles of graphs. Our investigation is concerned with the relationship between the families of graph ensembles, the structural features characteristic of these ensembles, and the sensitivity of the distances to these characteristic structural features.

The myriad proposed techniques for graph comparison [7] are severely reduced in number when one requires the practical restriction that the algorithm run in a reasonable amount of time on large graphs. Graph data frequently consists of 10^4 to 10^8 vertices, and so algorithms whose complexity scales quadratically with the size of the graph quickly become unfeasible. In this work, we restrict our attention to approaches where the calculation time scaled linearly or near-linearly with the number of vertices in the graph for sparse graphs.[4]

In the past 40 years, many random graph models have been developed which emulate certain features found in real-world graphs [12, 153]. A rigorous probabilistic study of the application of graph distances to these random models is difficult because the models are often defined in terms of a generative process rather than a distribution over the space of possible graphs. As such, researchers often restrict their attention to very small, deterministic graphs (see e.g. [108]) or to very simple random models, such as that proposed by Erdős and Rényi [49]. Even in these simple cases, rigorous probabilistic analysis can be prohibitively difficult. We adopt a numerical approach,

[4] Sparsity is, roughly, the requirement that the number of edges in a graph of size n be much lower than the maximum possible number $n^2/2$; a techincal definition is provided below.

in which we sample from random graph distributions and observe the empirical performance of various distance measures.

Throughout the work, we understand the observed results through a lens of global versus local graph structure. Examples of global structure include community structure and the existence of well-connected vertices (often referred to as "hubs"). Examples of local structure include the median degree in the graph, or the density of substructures such as triangles. Our results demonstrate that some distances are particularly tuned towards observing global structure, while some naturally observe both scales. In both our empirical and numerical experiments, we use this multi-scale interpretation to understand why the distances perform the way they do on a given model, or on given empirical graph data.

The chapter is structured as follows: in Section 5.2, we introduce the distances used, and establish the state of knowledge regarding each. In Section 5.3, we similarly introduce the random graph models of study and discuss their important features. In Section 5.4 we numerically examine the ability of the distances to distinguish between the various random graph models. The reader who is already familiar with the graph models and distances discussed can skip to Section 5.4.3 for a discussion of the results of our evaluation of the distances on the various random graph models, referencing the results in Section 5.4.2 as necessary. In Section 5.5, we apply the distances to empirical graph data and discuss the results. Finally, Section 5.6 summarizes the work and summarizes our recommendations. In the appendix, we introduce and discuss `NetComp`, the Python package written by the authors which implements the distances used to compare the graphs throughout the chapter.

5.2 Graph Distance Measures

Let us begin by introducing the distances we will use in this study, and discussing the state of the knowledge for each. We have chosen both standard and cutting-edge distances, with the requirement that the algorithms be computable in a reasonable amount of time on large, sparse graphs. In practice, this means that the distances must scale linearly or near-linearly in the size in

the graph.

We refer to these tools as "distance measures," as many of them do not satisfy the technical requirements of a metric. Although all are symmetric, they may fail one or more of the other requirements of a mathematical metric. This can be very problematic if one hopes to perform rigorous analysis on these distances, but in practice it is generally not significant. Consider the requirement of identity of indiscernibles, in which $d(G, G') = 0$ if and only if $G = G'$. we rarely encounter two graphs where $d(G, G') = 0$; we are more frequently concerned with an approximate form of this statement, in which we wish to deduce that G is similar to G' from the fact that $d(G, G')$ is small. Similarly, although the triangle inequality is foundational in approximation and proof methods in analysis, it is rarely employed in our process in applying these distances for anomaly detection.

5.2.1 Notation

We must first introduce the notation used throughout the chapter. It is standard wherever possible.

We denote by $G = (V, E, W)$ a graph with vertex set $V = \{1, \ldots, n\}$ and edge set $E \subseteq V \times V$. The function $W : E \to \boldsymbol{R}^+$ assigns each edge (i, j) in E a positive number, which we denote $w_{i,j}$. We call $n = |V|$ the **size** of the graph, and denote by $m \stackrel{\text{def}}{=} |E|$ the number of edges. For $i \in V$ and $j \in V$, we say $i \sim j$ if $(i, j) \in E$. The matrix $\boldsymbol{A}$ is called the **adjacency matrix**, and is defined as

$$A_{i,j} \stackrel{\text{def}}{=} \begin{cases} w_{i,j} & \text{if } i \sim j \\ 0 & \text{else} \end{cases}$$

The **degree** d_i of a vertex is defined as $d_i \stackrel{\text{def}}{=} \sum_{j \sim i} w_{i,j}$. The **degree matrix** $\boldsymbol{D}$ is the diagonal matrix of degrees, so $D_{i,i} = d_i$ and $D_{i,j} = 0$ for $i \neq j$. The **Laplacian matrix** (or just Laplacian) of G is given by $\boldsymbol{L} \stackrel{\text{def}}{=} \boldsymbol{D} - \boldsymbol{A}$. The **normalized Laplacian** is defined as $\boldsymbol{\mathcal{L}} \stackrel{\text{def}}{=} \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{D}^{-1/2}$, where the diagonal matrix $\boldsymbol{D}^{-1/2}$ is given by

$$D_{i,i}^{-1/2} \stackrel{\text{def}}{=} \begin{cases} 1/\sqrt{d_i} & \text{if } d_i \neq 0 \\ 0 & \text{else} \end{cases}$$

We refer to $\boldsymbol{A}$, $\boldsymbol{L}$, and $\boldsymbol{\mathcal{L}}$ as **matrix representations** of G. These are not the only useful matrix representation of a graph, although they are some of the most common. For a more diverse catalog of representations, see [159]. Note that other normalizations of the Laplacian matrix are possible; a popular choice is normalizing the rows so that they sum to one, which results in the transition matrix for a random walk on the graph. Our choice maintains symmetry, and thus strictly real eigenvalues and eigenvectors, a property which the row-normalized Laplacian lacks. A real spectrum simplifies computations, for example, when one wishes to form a basis of eigenvectors in order to decompose real-valued functions on the graph. However, the interpretability of the spectrum comes at the cost of interpretability of the matrix itself; while the row-stochastic normalization has an easy-to-understand function in terms of random walks, the interpretation of our normalized Laplacian $\boldsymbol{\mathcal{L}}$ is not so straightforward.

The **spectrum** of a matrix is the sorted sequence of eigenvalues. Whether the sequence is ascending or descending depends on the matrix in question. We denote the i^{th} eigenvalue of the adjacency matrix by $\lambda_i^{\boldsymbol{A}}$, where $\lambda_1^{\boldsymbol{A}} \geq \lambda_2^{\boldsymbol{A}} \geq \ldots \geq \lambda_n^{\boldsymbol{A}}$. The i^{th} eigenvalue of the Laplacian matrix are denotes by $\lambda_i^{\boldsymbol{L}}$, with the eigenvalues sorted in **ascending** order, so that $0 = \lambda_1^{\boldsymbol{L}} \leq \lambda_2^{\boldsymbol{L}} \leq \ldots \leq \lambda_n^{\boldsymbol{L}}$. We similarly denote the i^{th} eigenvalue of the normalized Laplacian by $\lambda_i^{\boldsymbol{\mathcal{L}}}$, with $\lambda_1^{\boldsymbol{\mathcal{L}}} \leq \lambda_2^{\boldsymbol{\mathcal{L}}} \leq \ldots \lambda_n^{\boldsymbol{\mathcal{L}}}$.

Two graphs are **isomorphic** if and only if there exists a map between their vertex sets under which the two edge sets are equal. Since our vertex sets are integers, we can simplify this definition. In particular, let us say that $G \cong G'$ if and only if there exists a permutation matrix $\boldsymbol{P}$ such that $\boldsymbol{A}' = \boldsymbol{P}^T \boldsymbol{A} \boldsymbol{P}$.

We say that a distance d requires **node correspondence** when there exist graphs G, G', and H such that $G \cong G'$ but $d(G, H) \neq d(G', H)$. Intuitively, a distance requires node correspondence when one must know some meaningful mapping between the vertex sets of the graphs under comparison.

5.2.2 Graph Distance Taxonomy

The distance functions we study divide naturally into two categories, which we will now describe. These categories are not exhaustive; many distance functions (including one we employ in our experiments) do not fit neatly into either category. Akoglu et al. [7] provide an alternative taxonomy; our taxonomy refines a particular group of methods they refer to as "feature-based".[5]

5.2.2.1 Spectral Distances

Let us first discuss spectral distances, also known as λ distances. We will briefly review the necessary background; for a good introduction to spectral methods of graph comparison, see [159].

We will first define the adjacency spectral distance; the Laplacian and normalized Laplacian spectral distances are defines similarly. Let G and G' be graphs of size n, with adjacency spectra $\lambda^{\boldsymbol{A}}$ and $\lambda^{\boldsymbol{A}'}$, respectively. The **adjacency spectral distance** between the two graphs is defined as

$$d_{\boldsymbol{A}}(G, G') \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{n} \left(\lambda_i^{\boldsymbol{A}} - \lambda_i^{\boldsymbol{A}'}\right)^2},$$

which is just the distance between the two spectra in the ℓ_2 metric. We could use any ℓ_p metric here, for $p \in [0, \infty]$. The choice of p is informed by how much one wishes to emphasize outliers; in the limiting case of $p = 0$, the metric returns the measure of the set over which the two vectors are different, and when $p = \infty$ only the largest element-wise difference between the two vectors is returned. Note that for $p < 1$ the ℓ_p distances are not true metrics (in paritcular, they fail the triangle inequality) but they still may provide valuable information. For a more detailed discussion on ℓ_p norms, see [127].

The Laplacian and normalized Laplacian spectral distances $d_{\boldsymbol{L}}$ and $d_{\boldsymbol{\mathcal{L}}}$ are defined in the exact same way. In general, one can define a spectral distance for any matrix representation of a graph; for results on more than just the three we analyze here, see [159]. Spectral distances are invariant under permutations of the vertex labels; that is to say, if $\boldsymbol{P}$ is a permutation matrix, then

[5] Note that the authors in [7] are classifying anomaly detection methods in particular, rather than graph comparison methods in general.

the spectrum of $\boldsymbol{A}$ is equal to the spectrum of $\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}$. This allows us to directly compare the topological similarity of two graphs without having to discover any mapping between the vertex sets.

In practice, it is often the case that only the first k eigenvalues are compared, where $k \ll n$. We refer to such truncated λ distances as λ_k distances. When using λ_k distances, it is important to keep in mind that the adjacency spectral distance compares the *largest* k eigenvalues, whereas the Laplacian spectral distances compare the *smallest* k eigenvalues. Comparison using the first k eigenvalues for small k allows one to focus on the community structure of the graph, while ignoring the local structure of the graph [90]. Inclusion of the higher-k eigenvalues allows one to discern local features as well as global. As we will see, this flexibility allows the user to target the particular scale at which they wish to examine the graph, and is a significant advantage of the spectral distances.

The three spectral distances used here are not true metrics. This is because there exist graphs G and G' that are cospectral but not isomorphic. That is to say, adjacency cospectrality occurs when $\lambda_i^{\boldsymbol{A}} = \lambda_i^{\boldsymbol{A}'}$ for all $i = 1, \ldots, n$, so $d_{\boldsymbol{A}}(G, G') = 0$, but $G \ncong G'$. Similar notions of cospectrality exist for all matrix representations; graphs that are cospectral with respect to one matrix representation are not necessarily cospectral with respect to other representations.

Little is known about cospectrality, save for some computational results on small graphs [61] and trees [159]. Schwenk proved that a sufficiently large tree nearly always has a cospectral counterpart [129]. This result was extended recently to include a wide variety of random trees [21]. However, results such as these are not of great import to us; the graphs examined are large enough that we do not enounter cospectrality in our numerical experiments. A more troubling failure mode of the spectral distances would be when the distance between two graphs is very small, but the two graphs have important topological distinctions. In Section 5.4.3, we will provide further insight into the effect of topological changes on the spectra of some of the random graph models we study.

The consideration above addresses the question of how local changes effect the overall spectral properties of a graph. Some limited computational studies have been done in this direction. For example, Farkas et al. [50] study the transition of the adjacency spectrum of a small world graph as

the disorder parameter increases. As one might expect, they observe the spectral density transition from a highly discontinuous density (which occurs when the disorder is zero, and so the graph is a ring-like lattice) to Wigner's famous semi-circular shape [154] (which occurs when the disorder is maximized, so that the graph is roughly equivalent to an uncorrelated random graph.)

From an analytical standpoint, certain results in random matrix theory inform our understanding of fluctuations of eigenvalues of the uncorrelated random graph (see Section 5.3 for a definition). These results hold asymptotically as we consider the k^{th} eigenvalue of a graph of size n, where $k = \alpha n$ for $\alpha \in (0, 1]$. In this case, O'Rourke [114] has shown that the the eigenvalue λ_k is asymptotically normal with asymptotic variance $\sigma^2(\lambda_k) = C(\alpha) \log n/n$. An expression for the constant $C(\alpha)$ is provided; see Remark 8 in [114] for the detailed statement of the theorem. This result can provide a heuristic for spectral fluctuations in some random graphs, but when the structure of these graphs diverges significantly from that of the uncorrelated random graph, then results such as these become less informative.

Another common question is that of interpretation of the spectrum of a given matrix representation of a graph.[6] How are we to understand the shape of the empirical distribution of eigenvalues? Can we interpret the eigenvalues which separate from this bulk in a meaningful way? The answer to this question depends, of course, on the matrix representation in question. Let us focus first on the Laplacian matrix $\boldsymbol{L}$, the interpretation of which is the clearest.

The first eigenvalue of $\boldsymbol{L}$ is always $\lambda_1^{\boldsymbol{L}} = 0$, with the eigenvector being the vector of all ones, $\mathbf{1} \in \mathbb{R}^n$. It is a well-known result that the multiplicity of the zero eigenvalue is the number of connected components of the graph, i.e. if $0 = \lambda_k^{\boldsymbol{L}} < \lambda_{k+1}^{\boldsymbol{L}}$, then there are precisely k connected components of the graph [35]. Furthermore, in such a case, the first k eigenvectors will indicate the components. In [90], an approximate version of this statement is made rigorous, in which the first k eigenvalues being small is an indicator of a graph being strongly partitioned into k clusters. This result justifies the use of the Laplacian in spectral clustering algorithms.

[6] "Spectral structure" might refer to the overall shape of the spectral density, or the value of individual eigenvalues separated from the bulk.

The eigenvalues of the Laplacian also have an interpretation analogous to the vibrational frequencies that arise as the eigenvalues of the continuous Laplacian operator ∇^2. To understand this analogy, consider the graph as embedded in a plane, with each vertex representing an oscillator of mass one and each edge a spring with elasticity one. Then, for small oscillations perpendicular to the plane, the Laplacian matrix is precisely the coupling matrix for this system, and so the eigenvalues give the square of the normal mode frequencies, $\omega_i = \sqrt{\lambda_i^{\boldsymbol{L}}}$. For a more thorough exposition of this interpretation of the Laplacian, see [53].

Maas [97] suggests a similar interpretation of the spectrum of the adjacency matrix $\boldsymbol{A}$. Consider the graph as a network of oscillators, embedded in a plane as previous. Additionally suppose that each vertex is connected to so many external non-moving points (by edges with elasticity one) so that the graph becomes regular with degree r. The frequencies of the normal modes of this structure then connect to the eigenvalues of $\boldsymbol{A}$ via $\omega_i = \sqrt{r - \lambda_i^{\boldsymbol{A}}}$.[7]

5.2.2.2 Matrix Distances

The second class of distances we will discuss are called **matrix distances**, and consist of direct comparison of the structure of pairwise affinities between vertices in a graph. These affinities are frequently organized into matrices, and the matrices can then be compared, often via an entry-wise ℓ_p norm.

We have discussed spectral methods for measuring distances *between* two graphs; to introduce the matrix distances, we will begin by focusing on methods for measuring distances *on graphs*; that is to say, the distance $\delta(v, w)$ between two vertices $v, w \in V$. Just a few examples of such distances include the shortest path distance [109], the effective graph resistance [48], and variations on random-walk distances [64]. Of those listed above, the shortest path distance is the oldest and the most thoroughly studied; in fact, it is so ubiquitous that "graph distance" is frequently used synonymously with shortest path distance [58].

[7] If the graph is already regular with degree d, then this interpretation is consistent with the previous, since the eigenvalues of $\boldsymbol{L} = d\boldsymbol{I} - \boldsymbol{A}$ are just $\lambda_i^{\boldsymbol{L}} = d - \lambda_i^{\boldsymbol{A}}$.

There are important differeces between the distances δ that we might choose. The shortest path distance considers only a single path between two vertices. In comparison, the effective graph resistance takes into account all possible paths between the vertices, and so measures not only the length, but the **robustness** of the communication between the vertices. This distinction is important when, for example, considering travel between two locations on a road network subject to high traffic.

How do these distances *on a graph* help us compute distances *between graphs*? Let us denote by $\delta : V \times V \to \mathbb{R}$ a generic distance on a graph. We need assume very little about this function, besides it being real-valued; in particular, it need not be symmetric, and we can even allow $\delta(v, v) \neq 0$.[8] Recalling that our vertices $v \in V = \{1, \ldots, n\}$ are labelled with natural numbers, we can then construct a matrix of pairwise distances $\boldsymbol{M}$ via $M_{i,j} \stackrel{\text{def}}{=} \delta(i, j)$.

The idea behind what we refer to as **matrix distances** is that this matrix $\boldsymbol{M}$ carries important structural information about the graph. Suppose that, for our given distance $\delta(\cdot, \cdot)$ graphs G and G' have corresponding matrices $\boldsymbol{M}$ and $\boldsymbol{M}'$. We can then compare G and G' via

$$d(G, G') = \|\boldsymbol{M} - \boldsymbol{M}'\| \tag{5.1}$$

where $\|\cdot\|$ is a norm we are free to choose.[9]

Let us elucidate a specific example of such a distance; in particular, we will show how the edit distance conforms to this description. Let $\delta(v, w)$ be defined as

$$\delta(v, w) = \begin{cases} 1 & \text{if } v \sim w \\ 0 & \text{else.} \end{cases} \tag{5.2}$$

Then the matrix $\boldsymbol{M}$ is just the adjacency matrix $\boldsymbol{A}$. If we use the norm

$$\|\boldsymbol{M}\| = \sum_{i,j=1}^{n} |M_{i,j}| \tag{5.3}$$

then we call the resulting distance $d(G, G') \stackrel{\text{def}}{=} \|\boldsymbol{A} - \boldsymbol{A}'\|$ the **edit distance**.

[8] When we say "distance" we implicitly assume that smaller values imply greater similarity; however, we can also carry out this approach with a similarity score, in which larger values imply greater similarity.

[9] We could use metrics, or even similarity functions here, although that may cause the function d to lose some desirable properties.

Of course, the usefulness of such a distance is directly dependent on how well the matrix $\boldsymbol{M}$ reflects the toplogical structure of the graph. The edit distance focuses by definition on local structure; it can only see changes at the level of edge perturbations. If significant volume changes are happening in the graph, then the edit distance will detect this, as do our other matrix distances. However, in our numerical experiments, we match the expected volume of the models under comparison, and so the edit distance is generally unable to discern between the models seen in Section 5.4.

We also implement the resistance-perturbation distance, first discussed in [108]. This distance takes the effective graph resistance $R(u, v)$, defined in [48], as the measure of vertex affinity. This results in a (symmetric) matrix of pairwise resistanced $\boldsymbol{R}$. The resistance-perturbation distance (or just resistance distance) is based on comparing these two matrices in the entry-wise ℓ_1 norm given in Equation (5.3).

The nice theoretical properties of the effective graph resistance [48] motivate our computational exploration of how well it reflects structure in realistic scenarios. Unlike the edit distance, the resistance distance is designed to detect global structural differences between graphs. A recent work [158] discusses the efficacy of the resistance distance in detecting community changes.

Finally, we look at DeltaCon, a distance based on the fast belief-propagation method of measuring node affinities [85]. To compare graphs, this method uses the fast belief-propagation matrix

$$\mathbf{S} \stackrel{\text{def}}{=} [\boldsymbol{I} + \epsilon^2 \boldsymbol{D} - \epsilon \boldsymbol{A}]^{-1} \tag{5.4}$$

and compares the two representations $\mathbf{S}$ and $\mathbf{S}'$ via the Matusita difference:

$$d(G, G') = \sqrt{\sum_{i,j} \left(\sqrt{S_{i,j}} - \sqrt{S'_{i,j}} \right)^2}. \tag{5.5}$$

Note that the matrix $\mathbf{S}$ can be rewritten in a matrix power series as

$$\mathbf{S} \approx \boldsymbol{I} + \epsilon \boldsymbol{A} + \epsilon^2 (\boldsymbol{A}^2 - \boldsymbol{D}) + \dots \tag{5.6}$$

and so takes into account the influence of neighboring vertices in a weighted manner, where neighbors separated by paths of length k have weight ϵ^k. Fast belief-propagation is designed to model the

diffusion of information throughout a graph [83], and so should in theory be able to perceive both global and local structures. Although empirical tests are performed in [85], no direct comparison to other modern methods are presented.

5.2.2.3 Other Graph Distances

These two categories do not cover all possible methods of graph comparison. The computer science literature explores various other methods (see [7], Section 3.2 for a nice review), and other disciplines that apply graph-based techniques often have their own idiosyncratic methods for comparing graphs extracted from data.

One possible method for comparing graphs is to look at specific "features" of the graph, such as the degree distribution, betweenness centrality distribution, diameter, number of triangles, number of k-cliques, etc. For graph features that are vector-valued (such as degree distribution) one might also consider the vector as an empirical distribution and take as graph features the sample moments (or quantiles, or statistical properties). A **feature-based distance** is a distance that uses comparison of such features to compare graphs.

Of course, in a general sense, all methods discussed so far are feature based; however, in the special case that the features occur as values over the space $V \times V$ of possible node pairings, we choose to refer to them more specifically as matrix distances. Similarly, if the feature in question is the spectrum of a particular matrix realization of the graph, we will call the method a spectral distance.

In [20], a feature-based distance called NetSimile is proposed, which focuses on local and egonet-based features (e.g. degree, volume of egonet as fraction of maximum possible volume, etc). If we are using k features, the method aggregates a feature-vertex matrix of size $k \times n$. This feature matrix is then reduced to a "signature vector" (a process they call "aggregation") which consists of the mean, median, standard deviation, skewness, and kurtosis of each feature. These signature vectors are then compared in order to obtain a measure of distance between graphs.

In the neuroscience literature in particular, feature-based methods for comparing graphs

are popular. In [146], the authors use graph features such as modularity, shortest path distance, clustering coefficient, and global efficiency to comopare functionaly connectivity networks of patiens with and without schizophrenia. Statistics of these features for the control and experiment groups are aggregated and compared using standard statistical techniques.

We implement NetSimile in our numerical tests as a prototypical feature-based method. It is worth noting that the general approach could be extended in almost any direction; any number of features could be used (which could take on scalar, vector, or matrix values) and the aggregation step can include or omit any number of summary statistics on the features, or can be ommited entirely. We implement the method as it is originally proposed, with the caveat that calculation of many of these features is not appropriate for large graphs, as they cannot be computed in linear or near-linear time. A scalable modification of NetSimile would utilize features that can be calculated (at least approximately) in linear or near-linear time.

5.2.3 Scaling and Complexity of Algorithms

In many interesting graph analysis scenarios, the sizes of the graphs to be analyzed are on the order of millions or even billions of vertices. For example, the social network defined by Facebook users has over 2 billion vertices as of 2017. In scenarios such as these, any algorithm of complexity $\mathcal{O}\left(n^2\right)$ will become unfeasible; although in principle it is possible that the constant M would be so small it would make up for the n^2 term in the complexity, in practice this is not the case. This motivates our requirement that our algorithms be of near-linear complexity. Indeed, even for graphs on the scale of 10^5, quadratic algorithms quickly become unfeasible.

This challenge motivates the previously stated requirement that all algorithms be of linear or near-linear complexity. We say an algorithm is **linear** if it is $\mathcal{O}(n)$; it is **near-linear** if it is $\mathcal{O}(n \log a_n)$ where a_n is asymptotically bounded by a polynomial. We use the notation $a_n = \mathcal{O}(b_n)$ in the standard way; for a more thorough discussion of algorithmic complexity, including definitions of the $\mathcal{O}$ notation, see [115].

We focus our attention on sparse graphs. We define sparsity as an asymptotic property, and

so it is only defined on a sequence of graphs. However, one can reasonably apply this to empirical graph data which changes over time and thus generates a natural time series which can be tested (roughly, since we are always at finite time) against this definition. In particular, let $\{G_i\}_{i=1}^{\infty}$ be a sequence of graphs, where the size of G_i is i and the number of edges in G_i denoted by m_i. We say a graph is sparse when the sequence $\{m_i\}_{i=1}^{\infty}$ is near-linear, in the sense given above.

Table 5.1 indicates the algorithmic complexity of each distance measure we compare. For DeltaCon and the resistance distance, there are approximate algorithms as well as exact algorithms; we list the complexity of both. Although we use the exact versions in our experiments, in practice the approximate version would likely be used if the graphs to be compared are large.

Of particular interest are the highly parallelizable randomized algorithms which can allow for extremely efficient matrix decomposition. In [63], the authors review many such algorithms, and discuss in particular their applicability to determining principal eigenvalues. The computation complexity in Table 5.1 for the spectral distances is based on their simplified analysis of the Krylov subspace methods, which states that the approach is $\mathcal{O}\left(kT_{\text{mult}} + (m+n)k^2\right)$, where T_{mult} is the cost of matrix-vector multiplication for the input matrix. Since our input matrices are sparse, $T_{\text{mult}} = \mathcal{O}(n)$, and $m + n = \mathcal{O}(n)$. Although we use the implicitly restarted Arnoldi method in our eigenvalue calculations, if implementing such a decomposition on large matrices the use of a randomized algorithm could lead to a significant increase in efficiency.

5.3 Random Graph Models

Random graph models have long been used as a method for understanding topological properties of graph data that occurs in the world. The uncorrelated random graph model of Erdős and Rényi [49] is the simplest model, and provides a null model akin to white noise. The tractability of this model has led to some beautiful probabilistic analysis [11] but the uniform topology of the model does not accurately model empirical graph data. The stochastic blockmodel is an extension of the uncorrelated random graph, but with explicit community structure reflected in the distribution over edges.

Distance Measure	**Complexity**	**Ref.**
Edit Distance	$\mathcal{O}(m)$	[10]
Resistance Distance (Exact)	$\mathcal{O}(n^2)$	[108]
Resistance Distance (Approximate)	$\mathcal{O}(m)$	[108]
DeltaCon (Exact)	$\mathcal{O}(n^2)$	[85]
DeltaCon (Approximate)	$\mathcal{O}(m)$	[85]
NetSimile	$\mathcal{O}(n \log n)$	[20]
Spectral Distance	$\mathcal{O}(nk^2)$	[63]

Table 5.1: Distance measures and complexity. n indicates the maximum of size of the two graphs being compares, and m indicates the maximum number of edges. For the spectral decomposition, k denotes the number of principal eigenvalues we wish to find. We assume that factors such as graph weights and quality of approximation are held constant, leading to simpler expressions here than appear in cited references. Spectral distances have equivalent complexity, since they all all amount to performing an eigendecomposition on a symmetric real matrix.

Models such as preferential attachment [12] and the Watts-Strogatz model [153] have been designed to mimic properties of observed graphs. Very little can be said about these models analytically, and thus much of what is understood about them is computational. The two-dimensional square lattice is a quintessential example of a highly structured and regular graph.

We will now introduce the models that are used. We study only undirected graphs, with no self-loops. Although directed graphs of are of great practical importance [164], tools such as the graph resistance only apply to undirected graphs. In particular, the electrical analogy needed to render the effective graph resistance menaingful is lost in a directed graps. Random-walk concepts are still perfectly meaningful on directed graphs, and motivate many popular algorithms used on such graphs (see e.g. [34]).

Most of the models in this work are sampled via the Python package NetworkX [62]; details of implementation can be found in the source code of the same. Some of the models we use are most clearly defined via their associated probability distribution, while others are best described by a generative mechanism. We will introduce the models roughly in order of complexity.

5.3.1 The Uncorrelated Random Graph

The **uncorrelated random graph** (also known as the Erdős-Rényi random graph) is a random graph in which each edge exists with probability p, independent of the existence of all others. We denote this distribution of graphs by $G(n, p)$ (recall that n denotes the size of the graph). As previously mentioned, this is by far the most thoroughly studied of random graph models; the simplicity of its definition allows for analytic tractability not found in many other models of interest. For example, the spectrum of the uncorrelated random graph is well understood. In particular, the spectral density forms a semi-circular shape, first described by Wigner [154], of radius $\sqrt{n}p$, albeit with a single eigenvalue $\lambda_1^{\boldsymbol{A}} \approx np$ separate from the semicircular bulk [50].

We will employ the uncorrelated random graph as our null model in many of our experiments. It is, in some sense, a "structureless" model; more specifically, the statistical properties of each edge and vertex in the graph are exactly the same. This model fails to produce many of properties

observed in empirical networks, which motivates the use of alternative graph models. For a more detailed definition of the model, and a thorough study of the properties of the uncorrelated random graph, see [11].

5.3.2 The Stochastic Blockmodel

One important property of empirical graphs is community structure. Vertices often form (relatively) densely connected communities, with the connection between communities being (relatively) sparse, or non-existent. This motivates the use of the **stochastic blockmodel**. In this model, each of the n vertices are in one two non-overlapping sets C_1 and C_2, referred to as "communities". Each edge $e = (i, j)$ exists (independently) with probability p if i and j are in the same community, and q if i and j are in separate communities. In this work, we will use "balanced" communities, so that the difference in size $|C_1| - |C_2|$ is less than 1 in magnitude.

The stochastic blockmodel is a prime example of a model which exhibits global structure without any meaningful local structure. In this case, the global structure is the partitioned nature of the graph as a whole. On a file scale, the graph looks like an uncorrelated random graph. We will use the model to determine which distances are most effective at discerning global (and in particular, community) structure.

The stochastic blockmodel is at the cutting edge of rigorous probabilistic analysis of random graphs. In particular, Abbe et al. [3] have recently proven a strict bound on community recovery, showing in exactly what regimes of p and q it is possible to discern the communities.

Generalizations of this model exist in which there are K communities of arbitrary size. Furthermore, each community need not have the same parameter p, and each community *pair* need not have the same parameter q. One can, in general, construct a (symmetric) matrix of parameters with the p_i on the diagonal, and the $q_{i,j}$ elsewhere, for $1 \leq i, j \leq K$. However, since our model has only two communities of nearly equal size, we only need a pair of parameters (p, q).

5.3.3 Preferential Attachment Models

Another often-studied feature of empirical graphs are their degree distribution. This is generally visualized as a histograph of degree frequencies, and studied under the assumption that it reflect some underlying distribution that informs us about the generative mechanism of the graph.

The degree distribution of an uncorrelated random graph is binomial, and so it has tails that exponentially decrease; for large d, the probability $\mathbb{P}(d)$ that a randomly chosen vertex has degree d decays exponentially, $\mathbb{P}(d) \propto \exp(-d^2)$. However, in observed graphs such as computer networks, human neural nets, and social networks, the observed degree distribution has a power-law tail [12]. In particular, one observed $\mathbb{P}(d) \propto d^{-\gamma}$ where generally $\gamma \in [2, 3]$. Such distributions are often also referred to as "scale-free".

The **preferential attachment model** is a scale-free random graph model. It is best described via the generative process rather than by a particular distribution over edges or possible graphs.[11] Although first described by Yule in 1925 [160], the model did not achieve its current popularity until the work of Barabási and Albert in 1999 [12].

The model has two parameters, l and n. The latter is the size of the graph, and the former controls the density of the graph. We require that $1 \leq l < n$. The generative procedure for sampling from this distribution proceeds as follows. Begin by initializing a star graph with $l + 1$ vertices, with vertex $l + 1$ having degree l and all others having degree 1. Then, for each $l + 1 < i \leq n$, add a vertex, and randomly attach it to l vertices already present in the graph, where the probability of i attaching to v is proportional to to the degree of v. Stop once the graph contains n vertices.

The constructive description of the algorithm does not yield itself to simpl analysis, and so less is known analytically about the preferential attachment model than the uncorrelated random graph or the stochastic blockmodel. In [51], the authors prove that the k^{th} eigenvalue of the Laplacian $\lambda_k^{\boldsymbol{L}}$ scales like $\lambda_k^{\boldsymbol{L}} \sim d_k^{1/2}$, where d_k is the k^{th} largest degree in the graph.[12] In [50], the authors demonstrate numerically that the adjacency spectrum exhibits a triangular peak with

[11] This feature of the preferential attachment model is what makes it particularly difficult to work with analytically.

[12] These results are proven on a model with a slightly different generative procedure; we do not find that they yield a particularly good approximation for our experiments, which are conducted at the quite low $n = 100$.

power-law tails.

Having a high degree makes a vertex more likely to attract more connections, so the graph quickly develops strongly connected "hubs," or vertices of high degree. This impacts both the global and local structure of the graph. Hubs are by definition global structures, as they touch a significant portion of the rest of the graph, making path lengths shorter and increasing connectivity throughout the graph. On the local scale, vertices in the graph tend to connect exclusively to the highest-degree vertices in the graph, rather than to one another, generating a tree-like topology. This particular topology yields a signature in the tail of the spectrum, the importance of which will be discussed below.

5.3.4 The Watts-Strogatz Model

Many real-world graphs exhibit the so-called "small world pheomenon," where the expected shortest path length between two vertices chosen uniformly at random grows logarithmically with the size of the graph. Watts and Strogatz [153] constructed a random graph model that exhibits this behavior, along with a high clustering coefficient not seen in an uncorrelated random graph. The clustering coefficient is defined as the ratio of number of triangles to the number of connected triplets of vertices in the graph. The **Watts-Strogatz model** [153] is designed to be the simplest random graph that has high local clustering and small average (shortest path) distance between vertices.

Like preferential attachment, this graph is most easily described via a generative mechanism. The algorithm proceeds as follows. Let n be the size of the desired graph, let $0 \leq p \leq 1$, and let k be an even integer, with $k < n$. We begin with a ring lattice, which is a graph where each vertex is attached to its k nearest neighbors, $k/2$ on each side. Then, for each edge in the graph (i, j) with $i < j$, with probability p rewire the edge to a random vertex l, so that (i, j) is replaced with (i, l). The target l is chosen so that $i \neq l$ and $i \not\sim l$ at the time of rewiring. Stop once all edges have been iterated through. In our implementations, we add an additional stipulation that the graph must be connected. If the algorithm terminates with a disconnected graph, then we restart the algorithm

and generate a new graph. This process is repeated until the resulting graph is connected.

As mentioned before, the toplogical features that are significant in this graph are the high local clustering and short expected distance between vertices. Of course, these quantities are dependent on the parameter p; as $p \to 1$, the Watts-Strogatz model approaches an uncorrelated random graph. Similarly, as $p \to 0$ the adjacency spectral density transitions from the tangle of sharp maxima typical of a ring-lattice graph to the smooth semi-circle of the uncorrelated random graph [50]. Unlike the models above, this model exhibits primarily local structure. Indeed, we will see that the most significant differences lie in the tail of the adjacency spectrum, which can be directly linked to the number of triangles in the graph [50]. On the large scale, however, this graph looks much like the uncorrelated random graph, in that it exhibits no communities or high-degree vertices.

This model fails to produce the scale-free behavior observed in many empirical graph data sets. Although the preferential attachment model reproduces this scale-free behavior, it fails to reproduce the high local clustering that is frequently observed, and so we should think of neither model as fully reaplicating the properties of observed graphs.

5.3.5 Random Degree-Distribution Graphs

The above three models are designed to mimic certain properties of empirical graphs. In some cases, however, we may wish to fully replicate a given degree sequence, while allowing other aspects of the graph to remain random. That is to say, we seek a distribution that assigns equal probability to each graph, conditioned upon the graph having a given degree sequence. The simplest model that attains this result is the configuration model [18]. Recently, Zhang et al. [161] have derived an asymptotic expression for the adjacency spectrum of a configuration model, which is exact in the limit of large graph size and large mean degree.

Inconveniently, this model is not guaranteed to generate a simple graph; the resulting graph can have self-edges, or multiple edges between two vertices. In 2010, Bayati et al. [17] described an algorithm which uniformly samples from the space of simple graphs with a given degree distribution.

We will refer to graphs sampled in this way as **random degree-sequence graphs**. Their utility lies in the fact that we can use them to control for the degree sequence when comparing graphs; they are used as a null model, similar to the uncorrelated random graph, but they can be tuned to share some structure (notably, the power-law degree distribution of preferential attachment) with the graphs to which they are compared.

The generative algorithm for this model is designed to sample from a uniform distribution over all possible graphs of a given size, conditioned upon the provided degree distribution. Their algorithm is fast, but not perfectly uniform; in [17] the authors prove that the distribution is asymptotically uniform, but do not prove results for finite graph size. We use this algorithm despite the fact that it does not sample the desired distribution in a truly uniform manner; the fact that the resulting graph is simple overcomes this drawback.

5.3.6 Lattice Graphs

In some of our experiments, we utilize lattice graphs. In particular we use a 2-dimensional n by m rectangular lattice. Using such a predictable structure allows us to test our understanding of our distances; in particular, we can see if our distances respond as we expect to structural features that are present in the lattice. Empirical realizations of planar graphs, such as road networks, often exhibit lattice-like structure. The planar structure of the lattice allows for an intuitive understanding of the spectral features of the graph, as they approximate the normal vibrational frequencies of a two-dimensional surface.

Lattice graphs are highly regular, in the sense that the connectivity pattern of each (interior) vertex is identical. This regularity is reflected by the discrete nature of the lattice's spectrum, which can be seen in Figure 5.10. This is a particularly strong flavor of local structure, as it is not subject to the nose present in random graph models. This aspect allows us to probe the functioning of our distances when they are exposed to graphs with a high amount of inherent structure and very low noise.

5.3.7 Exponential Random Graph Models

A popular random graph model is the **exponential random graph model**, or ERGM for short. Although they are popular and enjoy simple interpretability, we do not use ERGMs in our experiments. Unlike some of our other models which are described by their generative mechanisms, these are described directly via the probability of observing a given graph G.

In particular, let $g_i(G)$ be some scalar graph properties (e.g. size, volume, or number of triangles) and let θ_i be corresponding coefficients, for $i = 1, \ldots, K$. Then, the ERGM assigns to each graph a probability [95]

$$\mathbb{P}G \propto \exp\left(\sum_{i=1}^{K} \theta_i g_i(G)\right).$$

This distribution can be sampled via a Gibbs sampling technique, a process which is outlined in detail in [95]. ERGMs show great promise in terms of flexibility and interpretability; one can seemingly tune the distribution towards or away from any given graph metric, including mean clustering, average path length, or even decay of the degree distribution.

However, our experience attempting to utilize ERGMs led us away from this approach. When sampling from ERGMs, we were unable to control properties individually to our satisfaction. In particular, we found that attempts to increase the number of triangles in a graph increased the graph volume; when we subsequently used the ERGM parameters to *de-emphasize* graph volume, the sampled graphs had an empirical distribution very similar to an uncorrelated random graph.

5.4 Evaluation of Distances on Random Graph Ensembles

We will now present the results of our numerical tests, which compare the effectiveness of the various distances in discerning between pairs of random graph ensembles. The discussion in this section will be brief; we reserve our interpretation of the results until the next section. The experimens are organized via the models being compared, and with the performance of each distance shown in plots in each section. When appropriate, we also show the performance of the λ_k distances

for various k. Table 5.2 summarizes each comparison performed.

5.4.1 Description of Experiments

The experiments are designed to mimic a scenario in which a practicioner is attempting to determine whether a given graph belongs to a population or is an outlier relative to that population. In this vein, we perform experiments that determine how well each distance distinguishes populations drawn from a random graph model. In particular, let us define by $\mathcal{G}_0$ and $\mathcal{G}_1$ our two graph populations, which we will refer to as the **null** and **alternative** populations, respectively. For each distance measure, let $\mathcal{D}_0$ be the distribution of distances $d(G_0, G_0')$ where G_0 and G_0' are independently drawn from the distribution $\mathcal{G}_0$. Similarly, let $\mathcal{D}_1$ be the distribution of distances $d(G_0, G_1)$, where G_0 is drawn from $\mathcal{G}_0$ and G_1 is drawn from $\mathcal{G}_1$.[13]

The distances $\mathcal{D}_0$ are a characteristic distance between members of the population $\mathcal{G}_0$. The intuition here is that if the distribution of $\mathcal{D}_1$ is well separated from that of $\mathcal{D}_0$, then that distance is effective at separating the null population from the alternative population; if member of $\mathcal{G}_1$ are much further from members of $\mathcal{G}_0$ than this characteristic in-population distance, we can easily distinguish the two.

To that end, we normalize the statistics of $\mathcal{D}_1$ by those of $\mathcal{D}_0$ in order to compare. In particular, let μ_i be the sample mean of $\mathcal{D}_i$, and let σ_i be the sample standard deviation, for $i \in \{0, 1\}$. Then, we examine the statistics of $\widehat{\mathcal{D}}_1$, whose samples $\widehat{D}_1$ are calculated via

$$\widehat{D}_1 \stackrel{\text{def}}{=} \frac{D_1 - \mu_0}{\sigma_0}. \tag{5.7}$$

If our empirical distribution of $\widehat{\mathcal{D}}_1$ is well separated from zero (i.e. the mean is significantly greater than the standard deviation), then the distance is effectively separating the null and alternative population.

We generate 500 samples of D_0 and D_1, where each sample compares two graphs of size

[13] We use cursive to denote **distributions**, where as roman lettering denotes **samples** from those distributions, which are themselves random variables.

$n = 100$, unless otherwise specified. The graphs are always connected; our sampler will discard a draw from a random graph distribution if the resulting graph is disconnected. Said another way, we draw from the graph population conditioned upon the fact that the graph is connected. It is worth noting that, undercertain circumstances, this could significantly change the resulting graph population.

The small size of our graphs allows us to use larger sample sizes; although all of the matrix distances used have fast approximate algorithms available, we use the slower, often $\mathcal{O}\left(n^2\right)$, exact algorithms for our experiments, and so larger graphs would be prohibitivately slow to work with. In all the below experiments, we choose our parameters so that the expected volume $|E|$ of the two models under comparison is equal.

Sections 5.4.2.1 through 5.4.2.5 are separated by the models being compared. A very brief discussion of the results occurs alongside the presentation of the results, while a more thorough discussion is reserved for Section 5.4.3. The reader who wishes to primarly understand our observations and interpretation of the results can skip to 5.4.3 and reference the above sections as necessary.

5.4.2 Experimental Results

5.4.2.1 Stochastic Blockmodel

In Figure 5.1, we see the results of comparison between an uncorrelated random graph model and a stochastic block model. For the uncorrelated random graph, the probability of an edge existing is $p = 0.12$, which is a value of p for which the graph was almost always connected.[14]

For the stochastic blockmodel, we have two communities, each of size $n/2 = 50$, with parameters $(p, q) = (0.228, 0.012)$. Thus, the in-community connectivity is more dense than the cross-community connectivity by a factor of $p/q = 19$.

Since we have volume matched the graphs, the edit distance fails to distinguish the two

[14] In particular, with these parameters, we observe that the empirical probability of generating a disconnected uncorrelated random graph with these parameters is $\sim 0.02\%$. The preferential attachment section describes in more detail why this exact value is chosen.

models. Among the matrix distances, DeltaCon separates the two models most reliably. The adjacency and normalized Laplacian distances perform well, but the non-normalized Laplacian distance fails to distinguish the two models. The performance of the adjacency distance is primarily in the second eigenvalue $\lambda_2^{\boldsymbol{A}}$, and including further eigenvalues adds no benefit; the normalized Laplacian also shows most of its benefit in the second eigenvalue $\lambda_2^{\boldsymbol{\mathcal{L}}}$, but unlike the adjacency distance, including more eigenvalues decreases the performance of the metric.

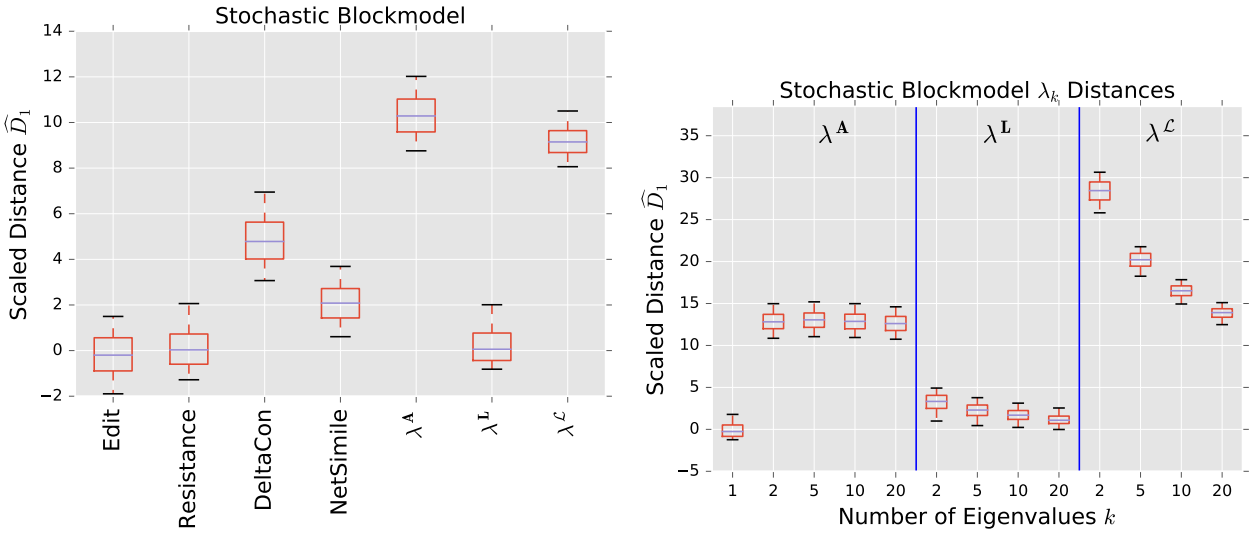

Figure 5.1: Comparison of distance performance, with uncorrelated random graph as null model and stochastic blockmodel as alternative. Boxes extend from lower to upper quartile, with center line at median. Whiskers extend from 5th to 95th percentile.

5.4.2.2 Preferential Attachment vs Uncorrelated

Figure 5.2 shows the results of comparing a preferential attachment graph to an uncorrelated random graph. The preferential attachment graph is quite dense, with $l = 6$. Since the number of edges in this model is always $|E| = l(n - l)$, we calculate the parameter p for the uncorrelated graph via

$$p(l) = l(n-l)\binom{n}{2}^{-1}.$$

We use $p(6) = 0.12$ in these experiments.

Again, due to matching the volumes of the graph, the edit distance fails to distinguish the two models. The resistance distance shows mediocre performance, although 0 is outside the 95% confidence interval. DeltaCon exhibits extremely high variability, although it has the highest median of the matrix distances.

The Laplacian distance outperforms all others, while the normalized Laplacian does not separate the two models at all. Figure 5.4 shows that most of the spectral information for the Laplacian is contained in the last few eigenvalues, counter to what one often expects from λ_k distances. For the adjacency distance, most of the information is held in the first eigenvalue, as the scaled distance stays more or less constant as one increases k (plot not shown).

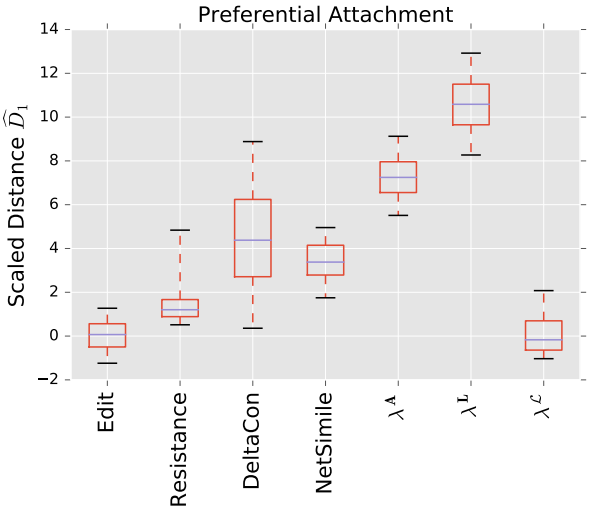

Figure 5.2: Comparison of distance performance, with uncorrelated random graph as null model and preferential attachment as alternative. See Figure 5.1 for boxplot details.

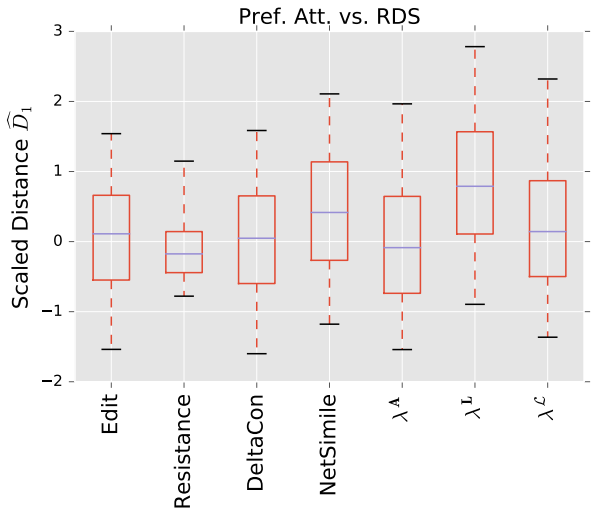

Figure 5.3: Comparison of distance performance, with degree-sequence random graph as null model and preferential attachment as alternative. The degree sequence for each null matches that of the alternative. See Figure 5.1 for boxplot details.

5.4.2.3 Preferential Attachment vs Random Degree Distribution Graph

In addition to the comparison of preferential attachment and uncorrelated random graphs in Section 5.4.2.2, we now compare preferential attachment to random degree-distribution graphs.

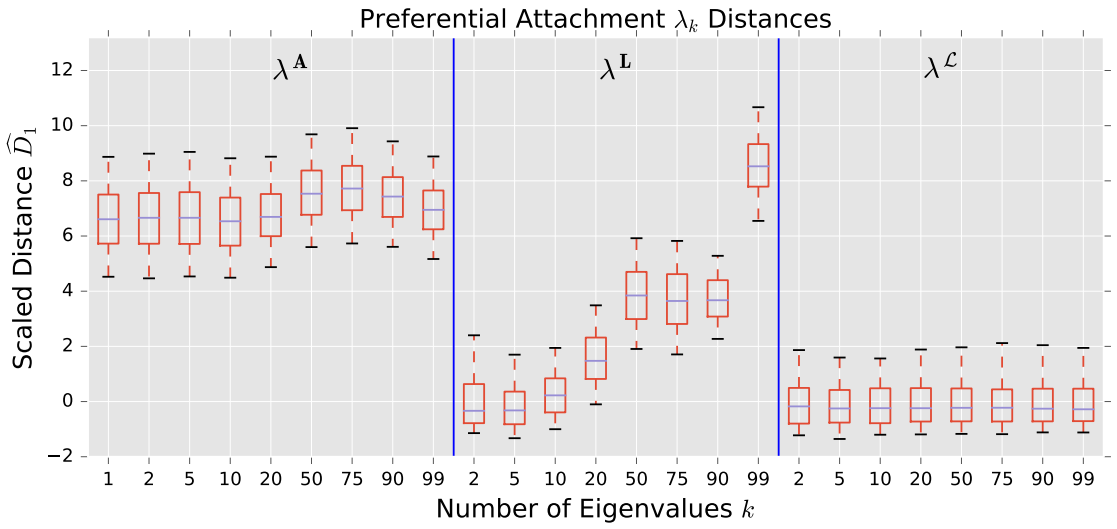

Figure 5.4: Comparison of λ_k distance performance, with uncorrelated random graph as null model and preferential attachment as alternative. See Figure 5.1 for boxplot details.

Recall that for a given degree distribution, the random degree-distribution graph probability density is the uniform density over all **simple** graphs with the given degree distribution. We employ the algorithm of Bayati et al. [17] to sample from this distribution.

This experiment allows us to search for structure in the preferential attachment model that is **not** perscribed by the degree distribution. The discrepancy in effectiveness of the normalized versus non-normalized Laplacian distances in Section 5.4.2.2 suggests that much of the structural information that the Laplacian distance is using to discern the two models is contained in the degree distribution. None of the metrics have scaled distances well-separated from zero, which suggests that all significant structural features of the preferential attachment model are perscribed by the degree distribution.

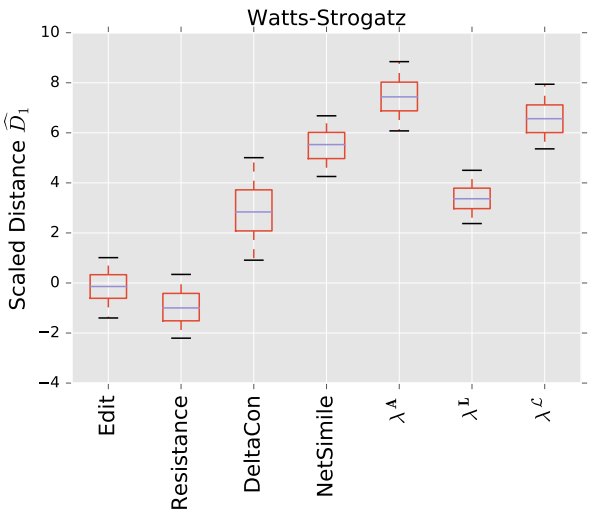

Figure 5.5: Comparison of distance performance, with uncorrelated random graph as the null model, and a small-world graph as the alternative. See Figure 5.1 for boxplot details.

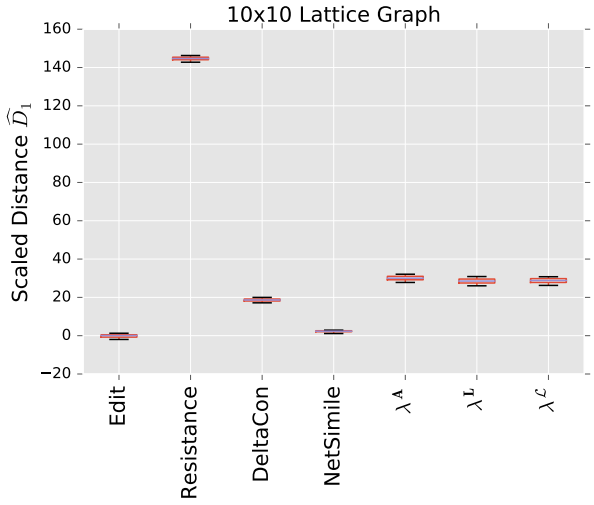

Figure 5.6: Comparison of distance performance, with a 10 by 10 2-dimensional lattice graph as the alternative model, and a random degree-distribution graph (with the same degree distribution as the lattice) as the null. See Figure 5.1 for boxplot details.

5.4.2.4 Watts-Strogatz

In this section, we compare a Watts-Strogatz random graph and an uncorrelated random graph. The Wattz-Strogatz model is interesting in that it contains primarily local structure, in the form of a high local clustering coefficient (i.e. density of triangles).

The Watts-Strogatz model is sparse, and so our volume-matched null model has a low value of p and thus is very likely disconnected. This is only a significant problem for the resistance distance, which is undefined for disconnected graphs. To remedy this, we use an extension of the resistance distance called the **renormalized resistance distance**, which is developed and analyzed in [158]. This is the only experiment in which the use of this particular variant of the resistance distance is required.

In Figure 5.5 we see that the adjacency and normalized Laplacian spectral distances are the strongest performers. Amongst the matrix distanced, DeltaCon strongly outperforms the resistance distance. The resistance distance here shows a negative median, which indicates smaller distances between populations than within the null population. This is likely due to the existence of many (randomly partitioned) disconnected components within this particular null model, which inflates the distances generated by the renormalized resistance distance. It is notable that, contrary to the comparison in Section 5.4.2.2, the normalized Laplacian outperforms the non-normalized version of the same.

In Figure 5.7 we see the results for λ_k distances, for a wide variety of k. These results indicate that much of the information that the λ distances are using to discern between the two models is contained in the higher eigenvalues, particularly for the adjacency and normalized Laplacian distances.

5.4.2.5 Lattice Graph

For our final experiment we compare a lattice graph to a random degree-distribution graph with the same degree distribution. The lattice here is highly structured, while the random degree-

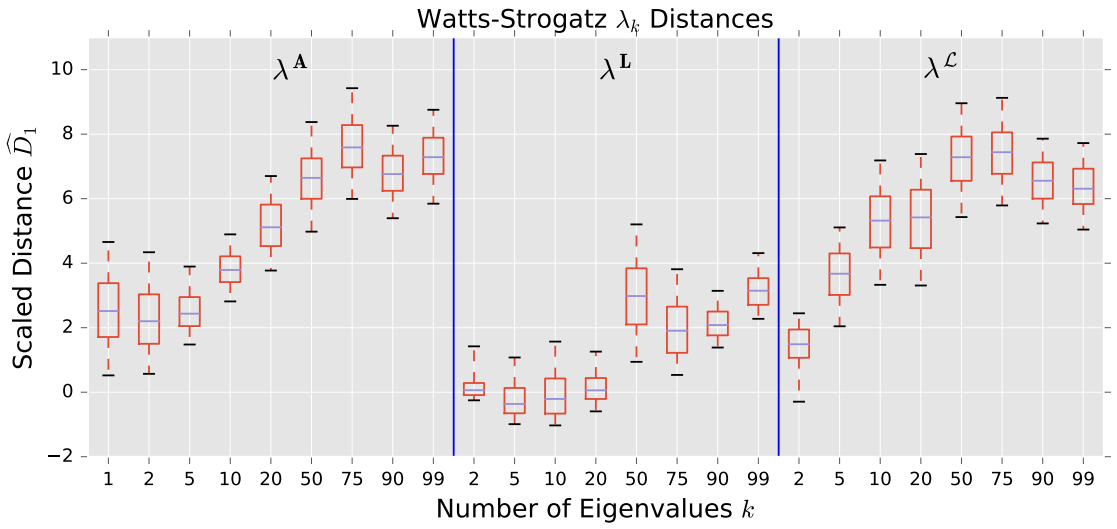

Figure 5.7: Comparison of λ_k distance performance, with uncorrelated random graph as the null model, and a small-world graph as the alternative. See Figure 5.1 for boxplot details.

distribution graph is quite similar to an uncorrelated random graph; both the deterministic degree distribution of the lattice and the binomial distribution of the uncorrelated random graph are highly concentrated around their mean.

We see that the scaled distances in this experiment are about an order of magnitude higher than they are in other experiments for some of the distances; because the lattice is such an extreme example of regularity, it is quite easy for many of the distances to discern between these two models. The resistance distance has the highest performance, while spectral distances all perform equally well. Note that for a regular graph, the eigenvalues of $\boldsymbol{A}$, $\boldsymbol{L}$, and $\boldsymbol{\mathcal{L}}$ are all equivalent, up to an overall scaling and shift, so we would expect near-identical performance for graphs that are nearly regular.

Similarly to the Watts-Strogatz comparison in Section 5.4.2.4, much of the information that the λ distances use to discern between the models is contained in the higher eigenvalues. This points to the importance of local structure in the lattice.

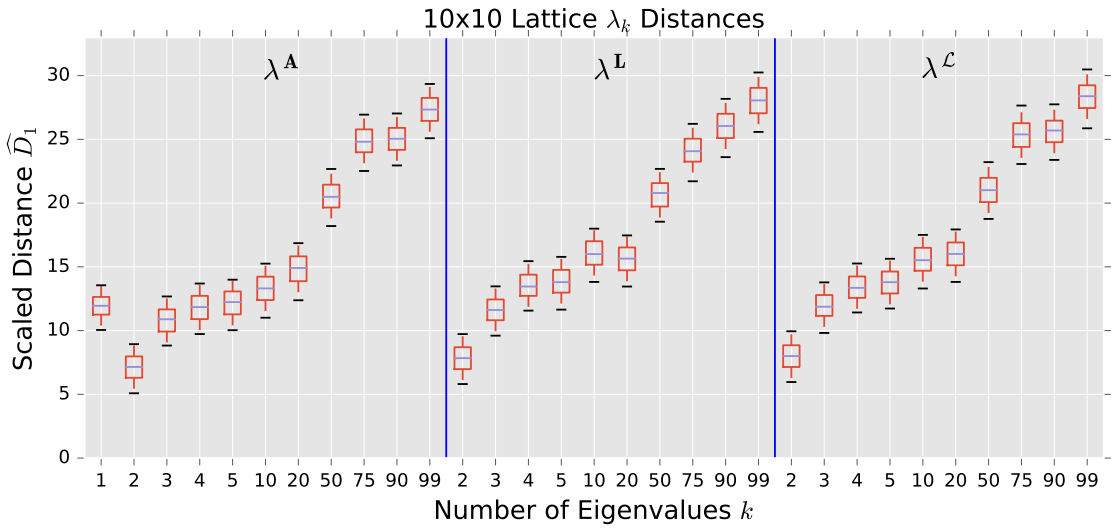

Figure 5.8: Comparison of λ_k distance performance, with a 10 by 10 2-dimensional lattice graph as the alternative model, and a random degree-distribution graph (with the same degree distribution as the lattice) as the null. See Figure 5.1 for boxplot details.

Section	Null	Alternative	Primary Structural Difference
5.4.2.1	$G(n,p)$	SBM	Community structure
5.4.2.2	$G(n,p)$	PA	High-degree vertices
5.4.2.3	RDDG	PA	Structure not in degree distr.
5.4.2.4	$G(n,p)$	WS	Local structure
5.4.2.5	$G(n,p)$	Lattice	Extreme local structure

Table 5.2: Table of comparisons performed, and the important structural features therein. $G(n,p)$ indicates the uncorrelated random graph, SBM is the stochastic blockmodel, PA is the preferential attachment model, RDDG is the random degree distribution graph, and WS is the Watts-Strogatz model.

5.4.3 Discussion

In this discussion, as we have done throughout the chapter, we will emphasize a distinction between local and global graph structure. Global structures include community separation as seen in the stochastic blockmodel, while local structures include the high density of triangles in the Watts-Strogatz model.

In general, we find that when examining global structure, the adjacency spectral distance and DeltaCon distance both provide good performance. When examining community structure in particular, one need not employ the full spectrum when using a spectral distance. The fact that the spectra of the graph provide a natural partitioning [90] aligns with our result that the first few eigenvalues will provide sufficient differentiation if the number of communities is low.

When one is interested in both global and local structure, we recommend use of the adjacency spectral distance. When the full spectrum is employed, the adjacency spectral distance is effective at differentiating between models even if the primary structural differences occur on the local level (e.g. the Watts-Strogatz graph). The use of the entire spectrum here is essential; much of the most important information is contained in the tail of the distribution, and the utility of the adjacency spectral distance decreases significantly when only the dominant eigenvalues are compared.

It is important to remember that these experiments represent only one way that pairwise graph comparison might be used. In particular, we are here comparing a sample to a known population. Alternatively, one might also be interested in comparing a dynamic graph at adjacent

timesteps; this scenario is treated empirically in Section 5.5.1.

5.4.3.1 Discerning Global Structure

Across our models, we see two significant and quite distinct types of global structure, which can be seen in Figure 5.9. The first of these is the grouping of the graph into distinct communities. The stochastic blockmodel is of course the model which most clearly possesses this type of global structure. At the local level, the stochastic blockmodel is nearly identical to the uncorrelated random graph, and so we can use the results of Section 5.4.2.1 to understand how distances respond to this specific feature.

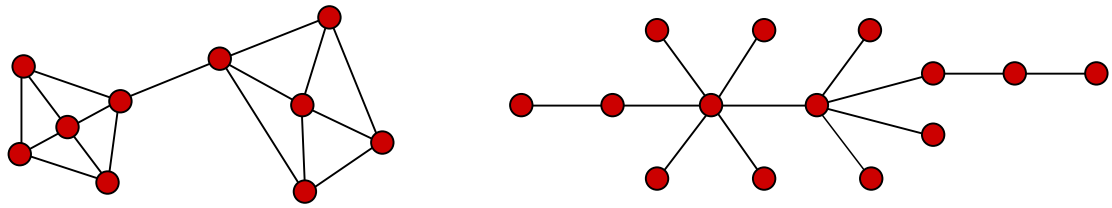

Figure 5.9: Two significant global structures observed in our experiments. On the left is the community structure typical of the stochastic blockmodel. On the right is the heavy-tailed degree distribution typical of the preferential attachment model.

The particular configuration of the stochastic blockmodel that we use has two partitions of equal size. We would thus expect the second eigenvalue λ_2 to be the primary distinguishing spectral feature of the graph (in any of the three matrix representations used). Indeed we see in Figure 5.1 that this is the case, and that the use of additional eigenvalues beyond $k = 2$ only serves to decrease performance by including noise in the comparison. In Figure 5.10 we can directly observe the similarity in spectra between the two models, as well as the presence in the stochastic blockmodel of a second eigenvalue which separates from the bulk of the spectrum.

The separation of the first k eigenvalues from the semi-circular bulk spectrum of the stochastic blockmodel is studied analytically in [161]. The authors show that a graph spectrum can be though of as two distinct components; a continuous bulk, and discrete outliers, with the latter indicating community structure. This separation is what allows our λ_k distances to function effectively in

detecting community structure. In general, the use of the spectrum for community partitioning in graphs has a long history [150]. Recent, Lee et al. [90] have proven a performance bound on the effectiveness of using the first k eigenvectors to partition the graph into k clusters.

In [158], the authors study the performance of the resistance perturbation metric in the setting of a dynamic variant of the stochastic blockmodel. Although not the same scenario as ours, their result is closely connected, and conforms well with the currently observed data. In particular, their result indicates that for the resistance metric to be effective in detecting topological changes in a stochastic blockmodel, the number of cross-community edges must be asymptotically dominated by the mean degree.

This is a highly restrictive condition. In the results shown in Section 5.4.2.1, we see that the resistance metrics performs poorly; auxilliary results (not shown) indicate that its performance increases significantly when the graph has high in-community degree $p = 0.495$ and very low cross-community connection $q = 0.005$. This unrealistic density requirement puts severe practical restrictions on the applicability of using the resistance metric to detect topological changes in community graphs. Furthermore, in these extreme cases, other measures (such as the spectral distances) can also easily distinguish between the two models.

The link between graph resistance and degree has been established in [152], where the authors show that the resistance R_{uv} between vertices u and v can be well approximated by

$$R_{uv} \approx \frac{1}{d_u} + \frac{1}{d_v},$$

an approximation which suggests that fluctuations in degree distribution would result in significant fluctuations in the graph resistance. This is corrobrated by the poor performance of the resistance distance in Section 5.4.2.1. These results indicate that the resistance distance cannot "see" changes in global structure over local noise, unless the global structure is unrealistically stark (as in the asymptotic condition given in [158]).

The second signifcant global structure seen in our models is the particular topology in which there are a small number of highly connected vertices which dominate the connectivity patterns

of the graph. A small graph exhibiting this structure can be seen in Figure 5.9. This results in a heavy-tailed degree distribution. The random graph model which features this type of structure is the preferential attachment model, whose degree distribution exhibits polynomial decay in the tails [12].

The best tool for detecting this structure is the Laplacian spectral distance. The presence of the degree matrix D in the Laplacian $L = D - A$ means that comparison of Laplacians is very effective for discerning between models with radically different degree distributions. Since signifcant differences between the degree distributions of the preferential and attachment graphs occur in the tail (i.e. high-degree vertices), the inclusion of the final few eigenvalues is essential if one wishes to use the Laplacian spectrum to perform this comparison.

Figure 5.10 exhibits the influence of the degree distribution on the Laplacian spectrum. We observe qualitatively, as demonstrated in [50], that the tails of the Laplacian spectrum of a preferential attachmend graph exhibits polynomial decay similar to the tail of the degree distribution. This is a prime example of the way in which the spectrum of the Laplacian can be heavily influenced by the degree distribution.

The particular topology of the preferential attachment differentiates itself from that of the uncorrelated random graph at both a global and local level. Even though the Laplacian spectral distance is best at observing the signifcant effect of high-degree vertices on the model, it is not, all in all, the most efficient tool for differentiating the two. To understand this, let us now turn to further discussion of the local structure present in the preferential attachment model, as well as the other models studied.

5.4.3.2 Impact of Local Structure

Local structure consists of structures existing at the level of a single vertex or subgraphs consisting of a small number of vertices. These local structures can provide important information about the topology of the graph, or they can amount to noise which obfuscates our ability to examine global structures of interest. Our experiments provide examples of both of these cases,

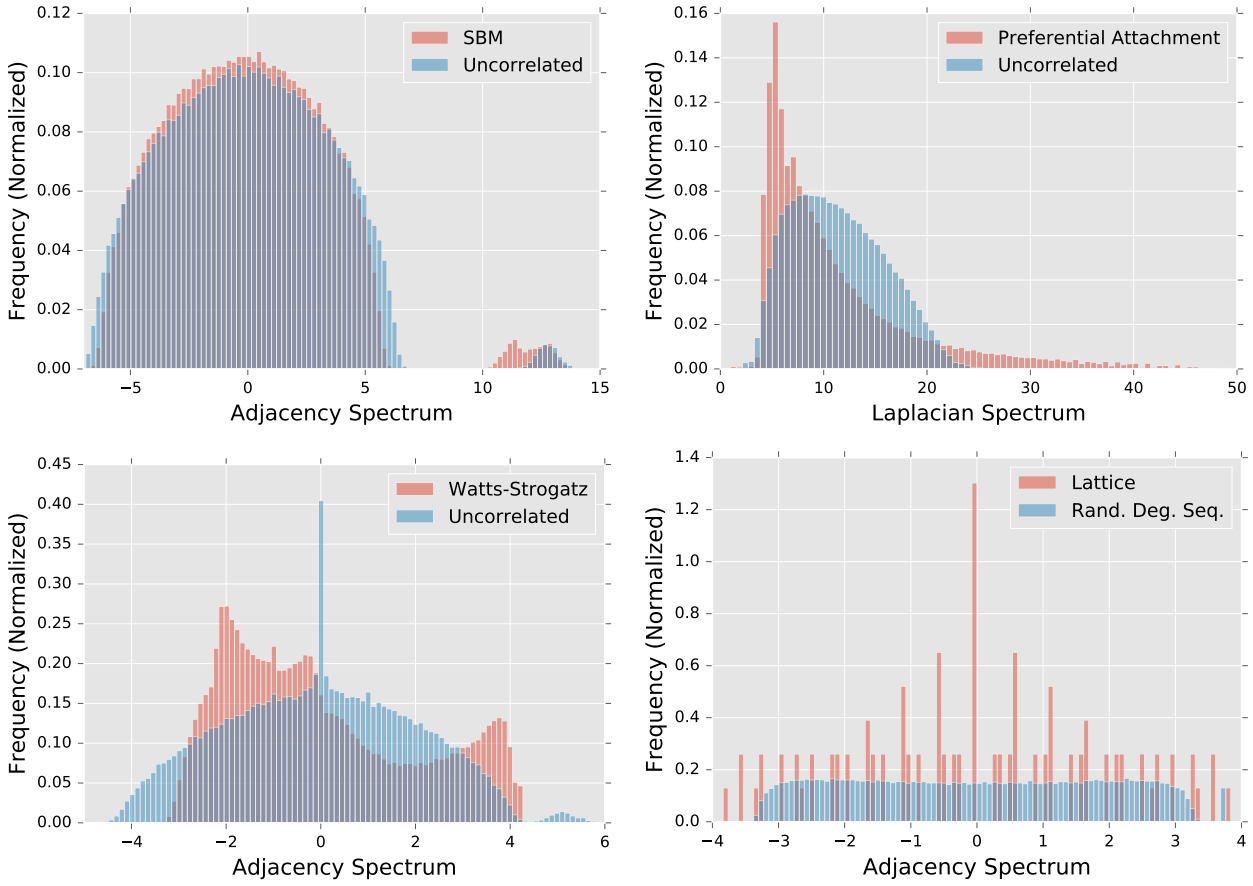

Figure 5.10: Spectral densities for various graph comparisons. Parameters used match those in Sections 5.4.2.1 through 5.4.2.5. Densities are built from an ensemble of 1000 graphs. The Laplacian spectrum $\lambda^{\boldsymbol{L}}$ is shown for preferential attachment, while adjacency spectrum $\lambda^{\boldsymbol{A}}$ is shown for all others. The uncorrelated random graph model in the lower left has a lower p than those on the upper row, resulting in a sharp peak at $\lambda^{\boldsymbol{A}} = 0$.

which we will now examine.

Consider first the results of Section 5.4.2.2. In Figure 5.4, we see that the adjacency spectral distance differentiates between the two models based primarily on the first dominant eigenvalue. Recalling the interpretation of the adjacency spectrum provided by Maas [97] and reiterated in Section 5.2.2.1, we realize that this is due to the high density of low-degree vertices in the preferential attachment model, compared to the uncorrelated random graph. This local structure is in some sense necessitated by the presence of a few very high degree vertices, since we demand that the graphs being compared are equal in expected volume. Indeed, the degree distribution of this model is so structurally significant that it almost entirely determines the structure of the resulting graph. We see in Figure 5.3 that no distances can effective discern between a preferential attachment graph and a randomized graph with the same degree distribution.

The Watts-Strogatz graph is another example of a model whose signature lies primarily in local structure. Farkas et al. [50] argue that the presence of a high number of triangles is the distinguishing feature of a Watts-Strogatz graph, and persists at values of β in which other structural aspects of the ring lattice (e.g. regularity and periodicity). The third moment of the spectral density of $\boldsymbol{A}$ tells us the expected number of triangles in a graph,[15] and so one would expect inclusion of the full spectrum important in detecting the toplogical signature of this model. On a global scale, the model does not signifcantly differ from the uncorrelated random graph; highly connected vertices are extremely unlikely, and the generative rewiring mechanism does not result in the presence of communities in the graph.

We see in Figure 5.7 that inclusion of the large-k (high frequency) eigenvalues is essential to differentiating between the models. In Figure 5.10 we see that the spectral density of the Watts-Strogatz model exhibits high skewness, which indicates the high expected number of triangles in the graph, and is only captured by inclusion of the full spectral bulk.

The lattice graph is an extreme example of this kind of local structure. Similarly to the

[15] This is not hard to show; see e.g. [50], Sec III A 1. Furthermore, the j^{th} moment of the density gives the expected number of paths of length j in the graph.

Watts-Strogatz model, there is a ubiquity of a certain type of local structure in the graph, namely the presence of four-edge loops. We see in Figure 5.8 that including a large number of eigenvalues in spectral comparison greatly increases the efficacy of the spectral distances. However, since the lattice is so remarkably regular (unlike the Watts-Strogatz model, which is a perturbed ring lattice) even comparing only a few principal eigenvalue is sufficient to differentiate it from a randomized graph with the same degree distribution. The spectra of the two models are shown in Figure 5.10.

Local structure is sometimes important when understanding graph structure, but can also frequently serve as a source of uninformative noise when comparing graphs. The results of Section 5.4.2.1 illustrate this fact. Looking to Figure 5.1, we see that the Laplacian spectrum is unable to distinguish between the stochastic blockmodel and the uncorrelated random graph, while the normalied Laplacian distinguishes them well. The difference between these two matrix representations is that normalization removes degree information, which is not informative in this particular model.

We see a similar problem arise when we apply the resistance distance to the stochastic blockmodel; as discussed in the previous section, the resistance distance is disproportionately influenced by local structure, and is unable to discern the global structure of the graph over local fluctuations. Interestingly, DeltaCon does not appear to suffer from local fluctuations as much as the resistance distance. This could be due to the structure of the matrix $\mathbf{S}$ that DeltaCon uses to represent the graph, or due to the use of the Matusita distance rather than the ℓ_1 or ℓ_2 norm to compare the resulting matrices (for more discussion of this, see Sections 2.2 and 3.1 in [85]).

It is essential to determine whether local toplogical features are of interest in the comparison problem at hand; inclusion of locally targeted distance measures can hinder the performance of graph distances in cases where local structure is noisy and uninformative. However, if local structure is ignored, one can often omit essential structural information about the graphs under comparison.

5.4.3.3 Recommendations

Throughout our experiments, the most consistent observation is that the adjacency spectral distance shows high effectiveness in discerning between a variety of models. We see in Section 5.4.2.1 that it is able to perceive global community structure and not be overwhelmed by local fluctuations in degree, but Sections 5.4.2.2 and 5.4.2.4 show that it is by no means ignorant of local structure present in a graph. That is to say, the adjacency spectral distance is **multiscale**; the scale of interest can be chosen by tuning the number of principal eigenvalues included in the comparison.

Spectral distances exhibits practical advantages over matrix distances, as they can inherently compare graphs of different sizes and can compare graphs without known vertex correspondence. The adjacency spectrum in particular is well-understood, and is perhaps the most frequently studied graph spectrum; see e.g. [50, 51]. Finally, fast, stable eigensolvers for symmetric matrices are ubiquitous in modern computing packages such as ARPACK, NumPy, and Matlab, allowing for rapid deployment of models based on spectral graph comparison.[16] Furthermore, randomized algorithms for matrix decomposition allow for highly parallelizable calculation of the spectra of large graphs [63].

However, the utility of the adjacency spectral distance is not general enough to simply apply it to any given graph matching or anomaly detection problem in a naive manner. A prudent practitioner would combine exploratory structural analysis of the graphs in question with an ensemble approach in which multiple distance measures are considered simultaneously, and the resulting information is combined to form a consensus. Such systems are commonplace in problems of classification in machine learning, where they are sometimes known as "voting classifiers" (see e.g. [123]).

As we have said before, we have been comparing graphs of equal volume (in expectation). In situations where the graph volume varies drastically, the process of chosing a graph comparison

[16] The Python library `NetComp` further simplifies the application of these tools to practical problems; see the appendix for more details.

tool may differ significantly. We will address this in Section 5.5.1, where we deal with graphs that exhibit significant volume fluctuations.

5.5 Applications to Empirical Data

Random graph models are often designed to simulate a single important feature of empirical networks, such as clustering in the Watts-Strogatz model or the high-degree vertices of the preferential attachment model. In empirical graphs, these factors coexist in an often unpredictable configuration, along with significant amounts of noise. Although the above analysis of the efficacy of various distances on random graph scenarios can help inform and guide our intuition, to truly understand their utility we must also look at how they perform when applied to empirical graph data.

In this section, we will examine the performance of our distance in two scenarios. First, we will look at an anomaly detection scenario for a dynamic social-contact graph, collected via RFID tags in an French primary school [138]. Secondly, we will look at a graph matching problem in neuroscience, comparing correlation graphs of brain activity in subjects with and without autism spectrum disorder [40].

The first experiment suggests that the tools that perform the most consistently in the graph matching applications (the spectral distances) are unreliable in our anomaly detection experiment. It is also interesting insofar as the graphs exhibit significant volume fluctuations, which was a factor not present in our numerical studies.

In the second experiment, we see that none of our graph distances fully distinguish between the two populations. Signal-to-noise is a ubiquitous problem in analyzing actual graph data, and is particularly notable in building a connectivity networks of human brain activity (see e.g. [30]). Accordingly, the results of our data experiments show that in the presence of real-world noise levels, many of these distances fail to distinguish subtle structural differences. In the face of this, we examine more targeted analysis techniques which may be applied in such a situation.

5.5.1 Primary School Social Contact Data

Some of the most well-known empirical network datasets reflect social connective structure between individuals, often in online social network platforms such as Facebook and Twitter. These networks exhibit structural features such as communities and highly connected vertices, and can undergo significant structural changes as they evolve in time. Examples of such structural changes include the merginf of communities, or the emergence of a single user as a connective hub between disparate regions of the graph.

In this section, we investigate a social contact network, which is based on measurements of face-to-face contact using RFID tags. We use our distances to compare the graph at subsequent timesteps. This is a quite different scenario than that presented in Section 5.4; the most important difference is that there is a natural sense of vertex correspondence, because the students' labels persist over time. This change has significant implications for the performance of our various distances, which we will explore in the discussion below.

5.5.1.1 Description of Experiment

The data are part of a study of face to face contact between primary school students [138]. Briefly, RFID tags were used to record face-to-face contact between students in a primary school in Lyon, France in October, 2009. Events punctuate the school day of the children (see Table 5.3), and lead to fundamental topological changes in the contact network (see Fig. 5.11). The school is composed of ten classes: each of the five grades (1 to 5) is divided into two classes (see Fig. 5.11).

Time	Event
10:30 a.m. – 11:00 a.m.	Morning Recess
12:00 p.m. – 1:00 p.m.	First Lunch Period
1:00 p.m. – 2:00 p.m.	Second Lunch Period
3:30 p.m. – 4:00 p.m.	Afternoon Recess

Table 5.3: Events that punctuate the school day.

The construction of a dynamic graph proceeds as follows: time series of edges that correspond to face to face contact describe the dynamics of the pairwise interactions between students. We divide the school day into $N = 150$ time intervals of $\Delta t \approx 200$ s. We denote by $t_i = 0, \Delta t, \ldots, (N-1)\Delta t$, the corresponding temporal grid. For each t_i we construct an undirected unweighted graph G_{t_i}, where the $n = 232$ nodes correspond to the 232 students in the 10 classes, and an edge is present between two students u and v if they were in contact (according to the RFID tags) during the time interval $[t_{i-1}, t_i)$.

Changes in the graph topology during the school day are quantified using the various distance measures,

$$D_R(t_i) \stackrel{\text{def}}{=} d(G_{t_{i-1}}, G_{t_i}).$$

To help compare these distances with one another, we normalize each by their sample mean $\overline{D} = N^{-1} \sum_i D(t_i)$, and we define

$$\widehat{D}(t) = D(t)/\overline{D}.$$

For the purpose of this work, we think of each class as a community of connected students; classes are weakly connected (e.g., see Fig. 5.11 at times 9:00 a.m., and 2:03 p.m.). During the school day, events such as lunchtime and recess, trigger significant increases in the the number of links between the communities, and disrupt the community structure; see Fig. 5.11 at times 11:57 a.m., and 1:46 p.m..

5.5.1.2 Discussion

In Figure 5.12, we see the normalized time series $\widehat{D}_t$ for each of the distance measures studied. Interestingly, the matrix distances all achieve passable performance, while NetSimile and the spectral distances are far too noisy to be of any practical use. As we see in Figure 5.11, the main structural changes that the graph undergoes are transitions into and out of a strong ten-community structure that reflects the classrooms of the school. For example, the adjacency matrix begins as (mostly) block-diagonal at 9 AM, but has signifcant off-diagonal elements by morning recess at

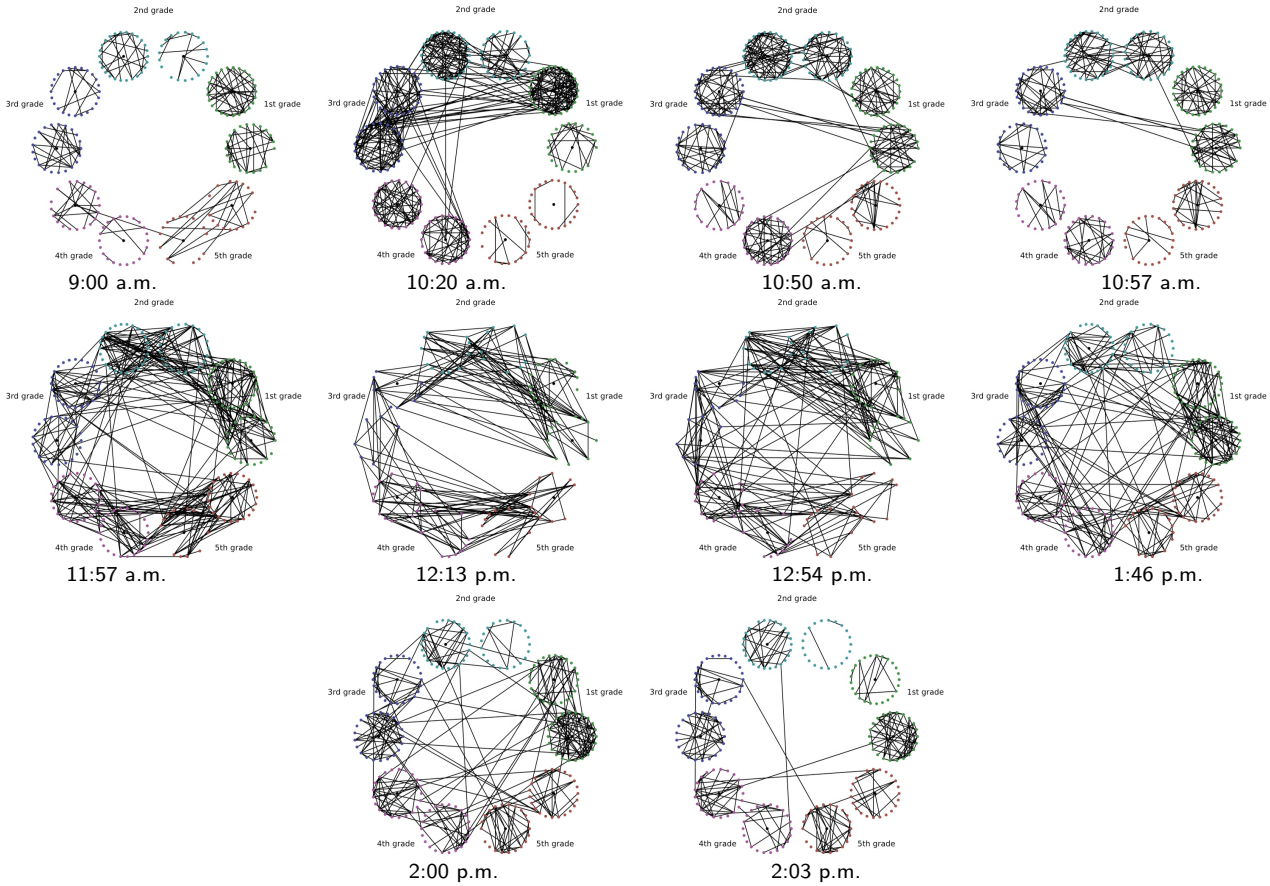

Figure 5.11: Top to bottom, left to right: snapshots of the face-to-face contact network at times (shown below each graph) surrounding significant topological changes.

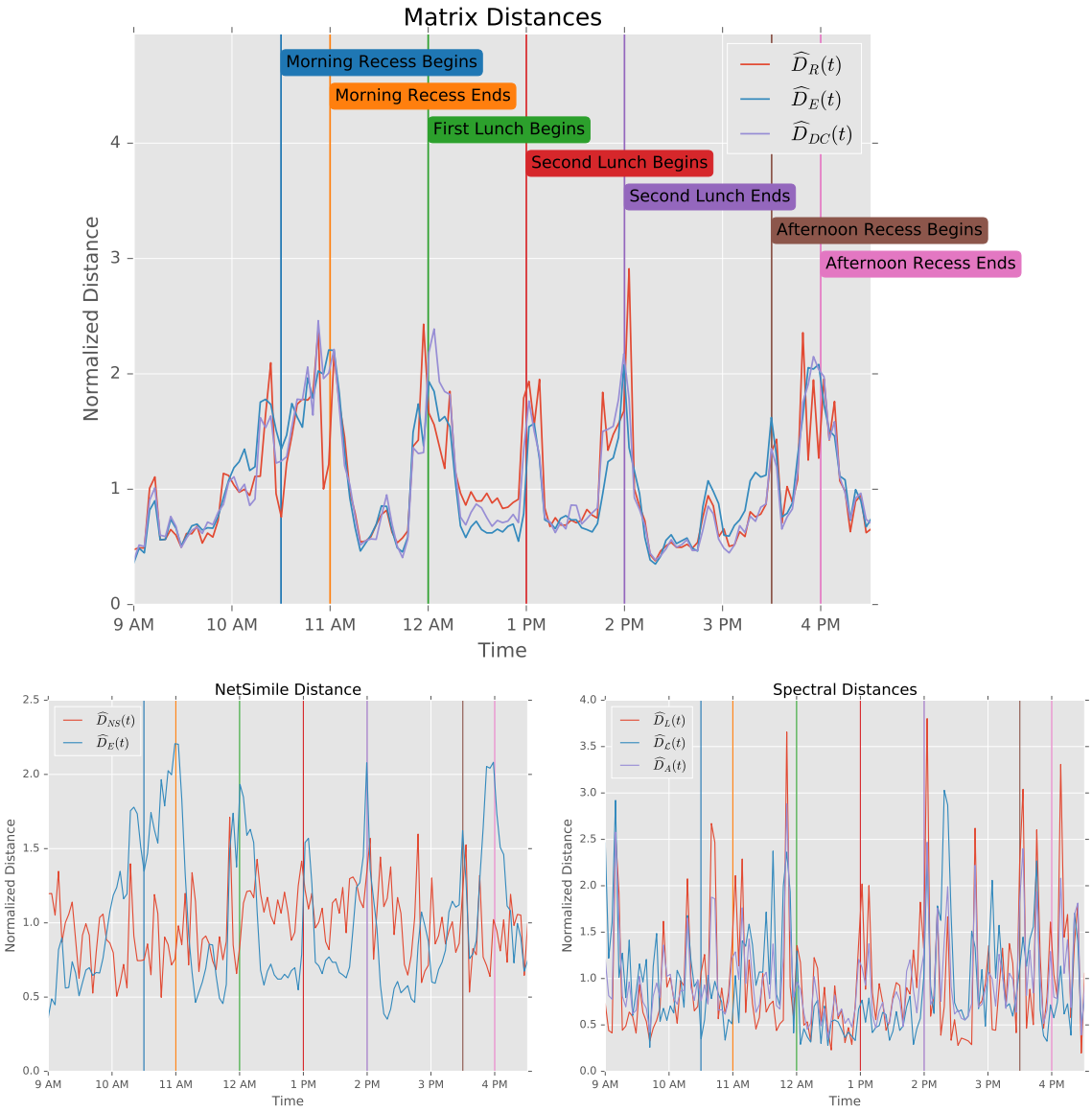

Figure 5.12: Comparison of distance performance on primary school data set. All distances are normalized by their sample mean.

10:20 AM, and is no longer (block) diagonally dominant come the lunch period at 12 PM.

These structural changes are of a global nature. In Section 5.4.2.1 we saw that the spectral distances were more effective than the matrix distances at detecting differences in community structure between graphs. Why is this not the case here? The graphs are persistent, in the sense that the vertices show a natural correspondence, which the matrix distances exploit. For example, we know that certain edges (those between classes) are "cross-community," and so the presence of these edges suggests some anomalous topology. This is why even the edit distance is quite effective at detecting topological changes in the graph.

Now let us highlight certain interesting features of the comparison shown in the top plot in Figure 5.12. Amongst the matrix distances, the resistance distance shows the largest distance at 10:20 AM, followed by the smallest distance at the subsequent timestep. During recess, local changes are occuring in the graph, but the global structure is remaining mostly constant - the graph consists of strongly connected communities, with some connection between them. The fact that the resistance distance shows the lowest distance between timesteps within recess suggests that, of the matrix distances, it is least effected by this local variation in the graph.

The lunch periods are marked by a stark change in graph topology. The graph undergoes significant global structural transformation, becoming almost entirely unordered with respect to class communities. Again the resistance distance shows a more signifcant signal at the beginning of this change, and a smaller signal as the anomalous topology persists.

The next significant transformation the graph undergoes is the transition out of lunch periods. The resistance distance stands out as marking this transition most pronouncedly, although all distances show significantly smaller distances between the time steps during the afternoon class period compared to during the lunch period.

Unlike our numerical experiments above, the graphs being compared here show significant fluctuations in volume. However, these fluctuations do not align with our event markers to the extent that our matrix distances do. This indicates that our matrix distances are picking up more than simply changes in volume.

The most remarkable conclusion of this particular experiment is that although the spectral distances are very efficient and stable for the purposes of graph matching, they show very poor performance in anomaly detection on dynamic graphs. This is due to the inherent vertex correspondence that is automatically provided when comparing a dynamic graph at subsequent timesteps. Although we reflect on the subtle distinctions in performance between the three matrix distances, they all show very similar overall performance, and any one of them would be sufficient for application in this scenario.

5.5.2 Brain Connectomics of Autism Spectrum Disorder

Graph theoretical analysis of the connective structure of the human brain is a popular research topic, and has benefitted from our growing ability to analyze network topology [29]. In these graph representations of the brain, the vertices are physical regions of the brain, and the edges indicate the connectivity between two regions. The connective structure of the brain is examined either at the "scructural" level, in which edges represent anatomical connection between two regions, or at the "functional" level, in which an edge connects regions whose activation patterns are in some sense similar.

Psychological conditions such as Alzheimer's disease [142], autism spectrum disorder [140], and schizophrenia [52] have been shown to have structural correlates in the graph representations of the brains of those affected. In this section, we will focus on autism spectrum disorder, or ASD. The availability of high-quality, open-access preprocessed data [38] makes ASD a particularly attractive choice for researchers with little experience implementing the nuanced preprocessing pipelines seen in neuroimaging. We examine which, if any, of our graph distances are able to effectively discern between subjects with ASD and subjects which are typically developing (TD). This is a problem in graph matching, very similar in structure to the experiments done on random graph models above.

We will see that our distances are ineffective at discerning between the graphs arising from ASD subjects and TD subjects. This result agrees with the conclusion of a recent review, which finds that classification methods do not generalize well to novel data [71]. The negative outcome of

this experiment both informs an understanding of the limitations of generalized tools such as our graph distances, and points to possible refinements of these tools.

5.5.2.1 Description of the Data

The Autism Brain Imagine Data Exachange, or ABIDE, is an aggregation of brain-imaging data sets from laboratories around the world which study the neurophisiology of ASD [40]. The data that we focus on are measurements of the activity level in various regions of the brain, measured via functional magnetic resonance imaging (fMRI). The fMRI method is actually measuring blood oxygen levels in the brain, which are then used as a proxy for activation levels. Measurements are taken over myriad small volumes within the brain, preprocessed, and then aggregated into a much smaller collection of timeseries, each representing a distinct region of the brain.

These timeseries then pass through an extensive preprocessing pipeline, which includes myriad steps such as nuisance signal (e.g. heartbeat) removal, detrending, smoothing via band-pass filtration, and so on. A detailed assessment of the preprocessing steps can be found in [38]. After preprocessing, the data is analyzed for quality. Of the original 1114 subjects (521 ASD and 593 TD), only 871 pass this quality-assurance step. These subjects are then spatially aggregated via the Automated Anatomical Labelling (AAL) atlas, which aggregates the spatial data into 116 time series.

To construct a graph from these time series, the pairwise Pearson correlation is calculated to measure similarity. If we let u and v denote two regions in the AAL atlas and let $P(u, v)$ denote the Pearson correlation between the corersponding time series, the simplest way to build a graph is to assign weights $w(u, v) = |P(u, v)|$, so the weight between vertices u and v is the modulus of the correlation. One may wish to exclude particularly low correlations, as these are often spurious and not informative as to the structure of the underlying network. In this case, one chooses a threshold T and assigns weights via

$$w(u,v) = \begin{cases} |P(u,v)| & |P(u,v)| \geq T \\ 0 & |P(u,v)| < T. \end{cases}$$

Finally, one may wish to binarize the graph, so that $w \in \{0, 1\}$, and thus uses the formula

$$w(u,v) = \begin{cases} 1 & |P(u,v)| \geq T \\ 0 & |P(u,v)| < T. \end{cases}$$

We will compare both binary and weighted connectomes, generated at multiple thresholding levels. This will allow us to be confident that our results are not artifacts of poorly chosen parameters in our definition of the connectome graph.

5.5.2.2 Discussion

Figure 5.13 shows the results of our experiment comparing connectomes of TD and ASD subjects. The TD subjects play the role of the null population $\mathcal{G}_1$ and the ASD subjects constitute the alternative population $\mathcal{G}_2$, and the scaled distances shown in the plot are calculated in the manner outlined in Section 5.4.

We observe that no distances effectively separate the two communities, regardless of level of thresholding, or the presence or absence of binarization. Indeed, the negative median of the scaled distance indicates that distances from ASD to TD connectomes is *lower* than distances between two TD connectomes, indicating a higher structural variability in TD connectomes. As we will see below, the structural differences between the two communities are localized within subgraphs, and do not persist throughout the full graph. Furthermore, the signal produced by these differences is not easily differentiated from the local variations (i.e. noise) present in the communities. For these two reasons, global comparison using graph metrics is ineffective for this problem.

Figure 5.14 shows a region-by-region comparison of connectomes of ASD and TD subjects. Similarly to our previous scalings, we take the mean and variance (region-wise) of the correlations in TD subject, and then use these to normalize the correlations of ASD subject. Thus, a value of 0.25 in Figure 5.14 indicates that ASD subject show, on average, correlations 0.25 standard deviations above the mean, relative to TD subjects.

We see that certain regions show significant differences. A closer examination shows that ASD subjects are generally underconnected in regions 73 through 77, and are overconnected in

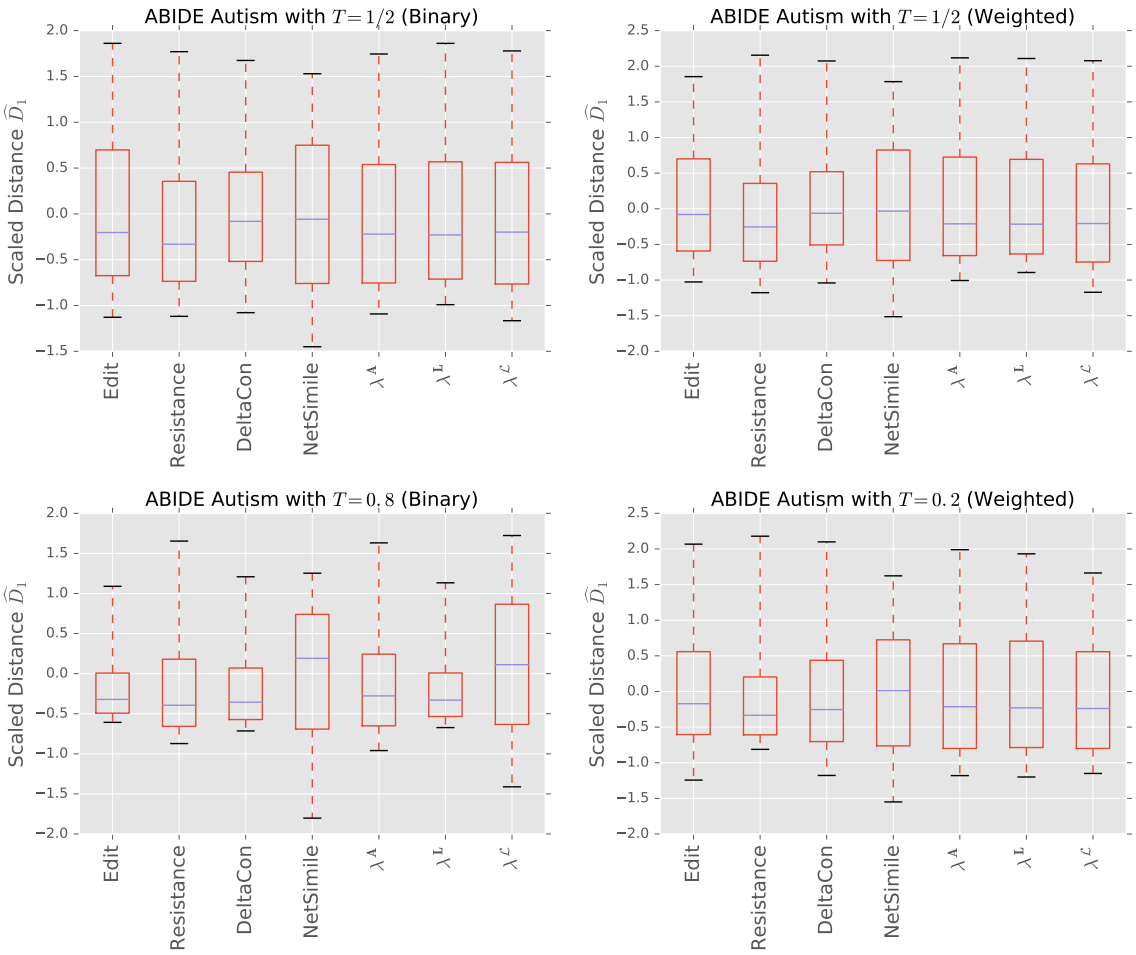

Figure 5.13: Distance comparison for ABIDE autism data set, for various thresholding configurations.

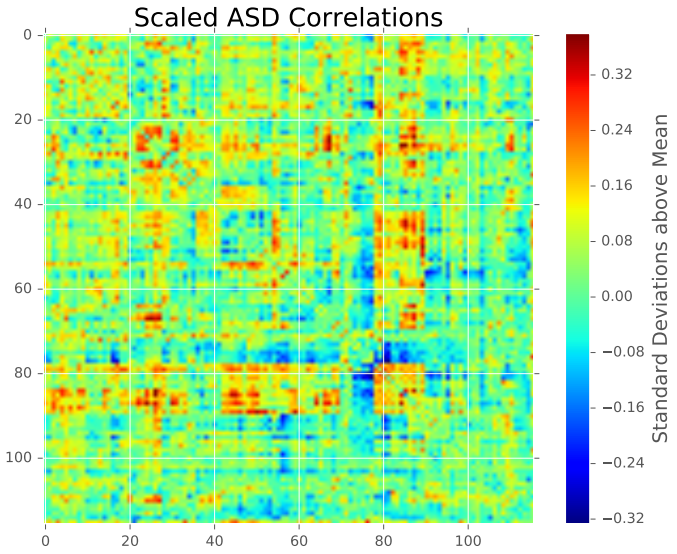

Figure 5.14: Absolute correlations between regions in ASD subjects. Data are normalized by mean and variance of absolute correlations between TD subjects.

regions 79 and 84 through 89. A table below shows the specific anatomical regions that these labels correspond to.

Label	Region	Connection
73	L. Putamen	Underconnected
74	R. Putamen	Underconnected
75	L. Globus Pallidus	Underconnected
76	R. Globus Pallidus	Underconnected
77	L. Thalamus	Underconnected
79	R. Transverse Temporal Gyrus	Overconnected
84	R. Superior Temporal Lobe	Overconnected
85	L. Middle Temporal Gyrus	Overconnected
86	R. Middle Temporal Gyrus	Overconnected
87	L. Middle Temporal Pole	Overconnected
88	R. Middle Temporal Pole	Overconnected
89	L. Inferior Temporal Gyrus	Overconnected

Table 5.4: Regions that show notably anomalous connectivity patterns. Correspondence between labels and regions is established via the Automated Anatomical Labelling atlas [145].

Figure 5.14 indicates that there are in fact significant structural differences between the connectomes of TD and ASD subjects. However, the differences barely stand out above the noise in the graph; all the edge differences in Figure 5.14 are less than half a standard deviation away from the mean. Furthermore, the differences occur in isolated regions of the graph, and the majority of the edge weights do not show statistically significant differences between the two populations. Both the low amplitude and small extent of the signal contribute to the difficulty we see graph distances have in discerning between TD and ASD subjects.

In [122], the authors find very little that distinguishes the connectomes of ASD subjects from TD subjects, save for lower betweenness centrality in right later parietal region. Similarly, the authors of [71] assert that although classification algorithms show "modest to conservatively good" accuracy rates, they perform poorly when tested on novel data sets. Looking in particular to [118], we see that the authors are able to achieve just over 75% accuracy in their classification, but they do so by both preprocessing the connectomes via particular regions of interest, and by testing a smorgasbord of classifiers (9 different models are used) to find the one which shows highest performance. Such a process exhibits that reasonable accuracy can be achieved via careful

algorithmic tuning; inversely, our result shows that a naïve application of a graph-theoretic method does not provide us with accurate classification.

A complete explanation of the failure of connectomes to unambiguously characterize ASD is well beyond the scope of this work; however, we will highlight a few interesting possibilities. In [118], the authors suggest that poor performance of many classifiers may be due to inclusion of uninformative features. Said another way, regions of the brain that have no bearing on the presence or absence of ASD are included in the connectome, which then reduces the signal to noise ratio of the data. In [71], the authors raise the issue of mental state in which the scans are performed. This state is often referred to as a "resting state," but variations in instructions (i.e. eyes open vs. closed) can have strong effects on the resulting data [143]. In [13], the authors find that when ASD connectomes show lower connectivity than TD connectomes when the subject exteroceptive activity, but show *higher* connectivity during introspective attentional tasks. This contrast highlights the need for careful control of the subjects' mental state if meaningful comparisons between connectomes are to be made.

It should also be noted that the choice of Pearson correlation to compare the fMRI time series is not obviously the correct one. It has been recently shown that the time series of brain activity exhibit nontrivial lag structure [104], indicating the need for a more general method of time series comparison. Myriad popular methods such as granger causality and mutual information can be applied to this problem; indeed, mutual information analysis has already shown promise as a tool in the pre-processing and feature selection stage of connectome analysis [103].

5.6 Conclusion

We have studied the efficacy of various graph distance measures when they are used to differentiate between popular random graph models, as well as in empirical anomaly detection and graph classification scenarios. These measures are understood through a multi-scale lens, in which the impact of global and local structures are considered separately. Although recent work [26] has called into question the previously assumed ubiquity of some of these models, studying their

properties builds qualitative and guide intuition when examining empirical network datasets.

Throughout our graph matching experiments in Section 5.4, we find that the adjacency spectral distance is the most stable, in the sense that it exhibits good performance across a variety of scenarios. It exhibits an ability to perceive both global and local structure, while avoiding becoming overwhelmed by local fluctuations in graph topology. Although the various matrix distances we examine allow for elegant analysis [85, 108], we find their performance on random graph model comparison underwhelming.

The situation reverses when we look at dynamic anomaly detection in Section 5.5.1. In this secario, the matrix distances proved most effective, and showed clear indications of the ground-truth anomalies present in the data. The spectral distances, on the other hand, were so noisy as to be useless. When doing anomaly detection on a dynamic graph, the two graphs in comparison tend to share more edges than in graph matching scenarios, which may contribute to the good performance of our matrix distances. Although the graphs in this section fluctuate in volume, we do not find that these fluctuations are helpful in detecting anomalous time steps.

Finally, we explore a collection of human connectomes of subjects with and without autism spectrum disorder. We observe that although differences are observable within the two populations via statistical comparison of edge weights, no graph distance effectively separates the two populations. This experiment helps us understand the limitations of using such generalized tools. We conclude that either more targeted tools are necessary, or more careful data collection and preprocessing is needed to establish a dataset that is separable via classification algorithms.

Based on the results of our numerical and empirical data experiments, we provide a suggested decision process in Figure 5.15. If the graphs to be compared exhibit differences in volume or size, then these should be examined to see if they hold predictive power, as they are so simple and efficient to compute. If they prove ineffective, then one must consider the setting. In a dynamic setting, in which a dynamic graph is being compared at subsequent timesteps, then we recommend using matrix distances based on the results of Section 5.5.1. If one is comparing graphs to determine whether a sample belongs to a given population, then the adjacency spectral distance

is the most reliable, as Section 5.4 demonstrates. Finally, if none of these approaches give adequate performance, then a more targeted analysis must be performed, such as the edge-wise statistical comparison of weights in Figure 5.14. The particular design of this analysis is domain specific and highly dependent upon the nature of the data.

5.7 Notation

For reference, in Table 5.5 we provide a table of notation used throughout the chapter.

5.8 NetComp: Network Comparison in Python

`NetComp` is a Python library which implements the graph distances studied in this work. Although many useful tools for network construction and analysis are available in the well-known `NetworkX` [62], advanced algorithms such as spectral comparisons and DeltaCon are not present. `NetComp` is designed to bridge this gap.

5.8.1 Design Consideration

The guiding principles behind the library are

(1) **Speed**. The library implements algorithms that run in linear or near-linear time, and are thus applicable to large graph data problems.[17]

(2) **Flexibility**. The library uses as its fundamental object the adjacency matrix. This matrix can be represented in either a dense (`NumPy` matrix) or sparse (`SciPy` sparse matrix) format. Using such a ubiquitous format as fundamental allows easy input of graph data from a wide variety of sources.

(3) **Extensibility**. The library is written so as to be easily extended by anyone wishing to do so. The included graph distances will hopefully be only the beginning of a full library of efficient modern graph comparison tools that will be implemented within `NetComp`.

[17] See below regarding the implementation of exact and approximate forms of DeltaCon and the resistance distance.

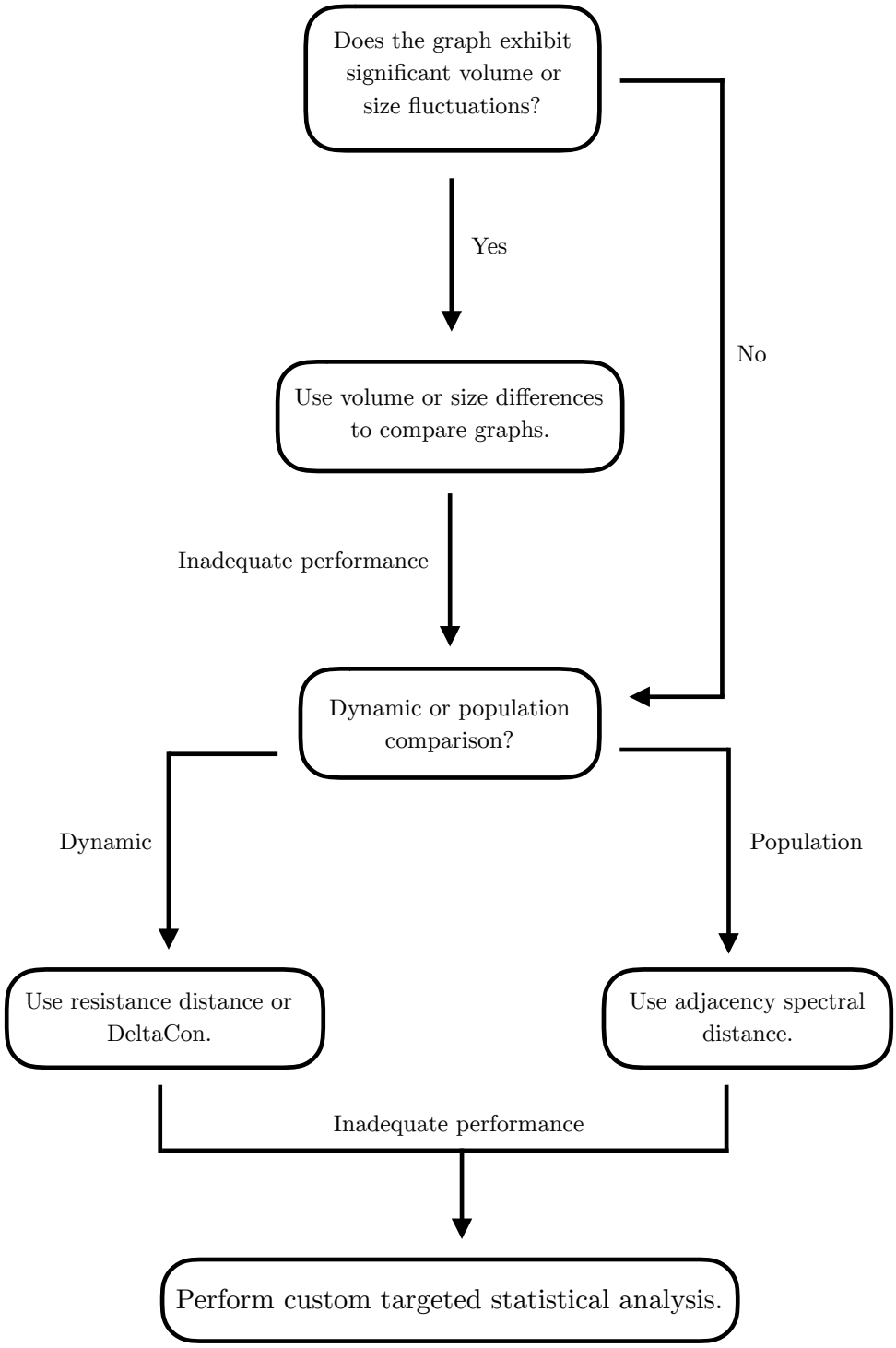

Figure 5.15: Flow chart summarizing the suggested decision process for applying distance measures in empirical data.

`NetComp` is available via the Python Package Index, which is most frequently accessed via the command-line tool `pip`. The user can install it locally via the shell command

`pip install netcomp`.

As of writing, the library is in alpha. The approximate (near-linear) forms of DeltaCon and the resistance distance are not yet included in the package. Both algorithms have an quadratic-time exact form which is implemeneted. Those interested can view the source code and contribute at `https://www.github.com/peterewills/netcomp`.

G	Graph
V	Vertex set, taken to be $\{1, 2, \ldots, n\}$
E	Edge set, subset of $V \times V$
W	Weight function, $W : E \to \mathbb{R}^+$
n	Size of the graph, $n = \lvert V \rvert$
m	Number of edges, $m = \lvert E \rvert$
d_i	Degree of vertex i
$\boldsymbol{D}$	Degree matrix (diagonal)
$d(\cdot, \cdot)$	Distance function
$\boldsymbol{A}$	Adjacency matrix
$\boldsymbol{L}$	Laplacian matrix
$\boldsymbol{\mathcal{L}}$	normalized Laplacian matrix (symmetric)
$\lambda_i^{\boldsymbol{A}}$	i^{th} eigenvalue of the adjacency matrix
$\lambda_i^{\boldsymbol{L}}$	i^{th} eigenvalue of the Laplacian matrix
$\lambda_i^{\boldsymbol{\mathcal{L}}}$	i^{th} eigenvalue of the normalized Laplacian matrix
$\mathcal{G}_{\{0,1\}}$	The {null,alternative} population of graphs
$G_{\{0,1\}}$	Sample graph from $\mathcal{G}_{\{0,1\}}$
$\mathcal{D}_0$	Distribution of distances between graphs in null population
D_0	Sample from $\mathcal{D}_0$
$\mathcal{D}_1$	Distribution of distances $d(G_0, G_1)$
D_1	Sample from $\mathcal{D}_1$
$\hat{\mathcal{D}}_1$	Distribution $\mathcal{D}_1$ normalized via (5.7)
$\hat{D}_1$	Sample from $\hat{\mathcal{D}}_1$
$G(n, p)$	Uncorrelated random graph with parameters n and p
(n, p, q)	Parameters for stochastic blockmodel
(n, l)	Parameters for preferential attachment model, with $1 < l \leq n$
(n, k)	Parameters for Watts-Strogatz graph, with $k < n$ even

Table 5.5: Table of commonly used notation.

Chapter 6

Conclusion

In Chapter 2, we have analyzed an approach based on test supermartingales [130] for performing hypothesis testing and constructing confidence intervals, and tested its performance in the scenario of estimating the success probability of Bernoulli trials. We showed that our test supermartingale method attains Bahadur efficiency, which is to say the resulting p-value has the highest possible exponential rate of decay. We also showed that the separation between the endpoint of a one-sided confidence interval and the true probability is $\mathcal{O}(\sqrt{\log n/n})$, while the optimal confidence interval endpoint has a separation of $\mathcal{O}(1/\sqrt{n})$. This separation can be though of as the cost of our method's robustness to stopping rules; we have proposed an adaptive modification which yields an optimal $\mathcal{O}(1/\sqrt{n})$ endpoint separation, provided that the number of trials is known in advance.

The law of the iterated logarithm suggests that our robustness to arbitrary stopping rules would incur a cost in endpoint separation of $\mathcal{O}(\log\log n)$, but our analysis shows a cost that is of order $\mathcal{O}(\log n)$ (under proper normalization). A natural extension of this work would be exploring efficient test supermartingales which, when inverted, generate confidence intervals with this optimal $\log\log n$ endpoint behavior. Furthermore, we might with to examine the effectiveness of the method when applied to trials which have more complicated distribution; for example, we might hope to study trials where the true mean or variance drifts over time, or in which subsequent trials show nontrivial dependence structure.

In Chapter 3, we studied the effects of thermal noise on a particular nanostructure arising in thin-film magnetics, called a magnetic droplet soliton, or simply a droplet. Through linearization of

this system around the droplet solution, we developed an analytical expression for the linewidth of the droplet, and the expected deviation of the droplet center. We find that the theoretical linewidth of the droplet agrees with micromagnetic simulations of the full (stochastic) partial differential system, particularly at large temperature. The center variability introduced by thermal noise is found to be orders of magnitude lower than that which is necessary to explain the imaging results found in [105], which suggests an alternative mechanism for droplet smearing during the imaging process. Our analysis has uncovered a previously unobserved regime of (deterministic) droplet instability, and has suggested drift instability as one possible mechanism for droplet smearing.

The timescales typical of droplet oscillation are deparated from the experimental timescaled in which droplets are observed by many orders of magnitude. Therefore, even if droplet ejection via thermal displacement is a very rare event, it may well occur $\sim 10^6$ times over the course of an experiment. A study of the statistics of rare events in our system would be necessary to provide a thorough understanding of the effect of thermal droplet ejection over experimental timescales. Furthermore, our approach has provided conclusions which are valid within certain experimental regimes, namely large nanocontact radii ρ_*. It remains to be seen how the effects of thermal noise impact systems in which the nanocontact is small, but one would expect reduced stability and thus an increase in the impact of thermal noise.

In Chapter 4, we studied the applicability of the effective graph resistance [48] to the detection of transitions in community structure in the dynamic stochastic blockmodel. We proved that, with high probability, the effective resistance will discern such transitions when the number of cross-community connections in the model is dominated by the average in-community degree. The crux of the proof of this result lies in the development of an expression for the effective resistance between arbitrary vertices in the stochastic blockmodel. This result analytically establishes the domain of applicability of the effective resistance when applied to graphs undergoing changes in community structure.

Most graphs with community structure posess more than two communities, and also have vertices which are not discernibly members of any of the communities present in the graph. A

natural extension of the work in Chapter 4 is to a stochastic blockmodel with more than two communities, communities of varying sizes, and with various configurations of intra-community connectivity. Furthermore, our work prove that a hypothesis test constructed from the resistance-perturbation distance has asymptotic power of 1, it does not explicitly establish the significance level of this hypothesis test. An explicit expression for the significance for finite graph size would inform the applicability of this approach to empirical data.

Chapter 5 has extended this study to a thorough analysis of distance measures for pairwise graph comparison, with a focus on distances which can be computed efficiently, and so can be applied to large graph datasets. Through both numerical studies of random graph models and analysis of empirical graph data, we have shown that the choice of the distance depends crucially on both the scale of structure one wishes to study as well as the scenario in which the analysis is being performed. In particular, the methods that are effective for discerning distinct graphs belonging to a given population are quite different from those which are effective for studying a graph evolving in time. We have provided a decision protocol for determining the most useful graph distance tool for a given application.

This work has highlighted the ability of various distances to capture various typical structural features of a graph, such as community structure or clustering. However, the form in which these structures appear in empirical graph data is highly varied, and they are far from the only important aspects. To truly understand the performances of these distances "in the wild," a thorough study of tens or hundreds of empirical graph data sets is necessary.

An additional challenge is that many graphs do not conform to our general prescription of simplicity. Recall that a simple graph is undirected, has no self-loops, and has no multi-edges. Furthermore, this study focuses on unweighted graphs; many empirical graph data sets are weighted. Although we have analyzed one empirical data set which includes weighted graphs in Section 5.5.2, we still lack a structural understanding of weighted graphs akin to the perspective that Section 5.4.3 offers on simple graphs.

It is worth noting that the extension of some of these methods to the directed graph setting is

not trivial; for example, the nature of the graph spectrum changes fundamentally when the matrix representations are asymmetric, and the effective graph resistance is not well-defined for directed graphs. Proposing and analyzing extensions of these distances that can apply to asymmetric graphs would also have great value.

Finally, although we have focused on graph distances which can be computed efficiently, the literature lacks a thorough study of the computational speed of these algorithms for finite graph size. One would hope to see the speed of computation compared for graphs of different scales and sparsities. Most results on the computational efficiency of algorithms is focused on asymptotic statements of complexity, which can obfuscate the actual finite-time computational cost, and do not take into account issues such as parallelizability (see [63]).

Bibliography

[1] The autism brain imaging data exchange. `http://fcon_1000.projects.nitrc.org/indi/abide/`.

[2] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. IEEE Transactions on Information Theory, 62(1):471–487, 2016.

[3] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. IEEE Transactions on Information Theory, 62(1):471–487, 2016.

[4] A. Acín and L. Masanes. Certified randomness in quantum physics. Nature, 540:213–219, 2016.

[5] Richard M Adler. A dynamic social network software platform for counter-terrorism decision support. In Intelligence and Security Informatics, 2007 IEEE, pages 47–54. IEEE, 2007.

[6] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery, 29(3):626–688, 2014.

[7] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery, 29(3):626–688, May 2015.

[8] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. Rev. Modern Phys., 74:47–97, 2002.

[9] D. Backes, F. Macià, S. Bonetti, R. Kukreja, H. Ohldag, and A. D. Kent. Direct observation of a localized magnetic soliton in a spin-transfer nanocontact. Phys. Rev. Lett., 115:127205, Sep 2015.

[10] Brian Baingana and Georgios B Giannakis. Joint community and anomaly tracking in dynamic networks. IEEE Transactions on Signal Processing, 64(8):2013–2025, 2016.

[11] B Ballobás. Random Graphs. Cambridge University Press, 2001.

[12] Albert-László Barabási and Réka" Albert. Emergence of scaling in random networks. Science, 286(5439):509–512, October 1999.

[13] Pablo Barttfeld, Bruno Wicker, Sebastián Cukier, Silvana Navarta, Sergio Lew, Ramón Leiguarda, and Mariano Sigman. State-dependent changes of connectivity patterns and functional brain network topology in autism spectrum disorder. Neuropsychologia, 50(14):3653–3662, 2012.

[14] V. G. Baryakhtar, B. A. Ivanov, A. L. Sukstanskii, and E. Yu. Melikhov. Soliton relaxation in magnets. Phys. Rev. B, 56:619–635, Jul 1997.

[15] Juan I Fuxman Bass, Alos Diallo, Justin Nelson, Juan M Soto, Chad L Myers, and Albertha JM Walhout. Using networks to measure similarity between genes: association index selection. Nature methods, 10(12):1169–1176, 2013.

[16] Michael Baur and Marc Benkert. Network comparison. In Network analysis, pages 318–340. Springer, 2005.

[17] Mohsen Bayati, Jeong Han Kim, and Amin Saberi. A sequential algorithm for generating random graphs. Algorithmica, 58(4):860–910, December 2010.

[18] E A. Bender and E. R. Canfield. The asymptotic number of labeled graphs with given degree sequences. Journal of Combinatorial Theory, Series A, pages 296–307, 1978.

[19] L. Berger. Emission of spin waves by a magnetic multilayer traversed by a current. Phys. Rev. B, 54(13):9353–9358, Oct 1996.

[20] Michele Berlingerio, Danai Koutra, Tina Eliassi-Rad, and Christos Faloutsos. Netsimile: A scalable approach to size-independent network similarity. CoRR, abs/1209.2684, 2012.

[21] Shankar Bhamidi, Steven N. Evans, and Arnab Sen. Spectra of large random trees. Journal of Theoretical Probability, 25(3):613–654, 2012.

[22] P. Bierhorst. A robust mathematical model for a loophole-free Clauser-Horne experiment. J. Phys. A, 48:195302, 2015.

[23] L. D. Bookman and M. A. Hoefer. Analytical theory of modulated magnetic solitons. Phys. Rev. B, 88:184401, Nov 2013.

[24] L. D. Bookman and M. A. Hoefer. Perturbation theory for propagating magnetic droplet solitons. Proceedings of the Royal Society of London A, 471(2179), 2015.

[25] P. J. Brockwell and R. A. Davis. Time Series: Theory and Methods. Springer Verlag, New York, NY, 2nd edition, 1991.

[26] Anna D. Broido and Aaron Clauset. Scale-free networks are rare. Submitted, 01 2018.

[27] William Fuller Brown. Thermal fluctuations of a single-domain particle. Phys. Rev., 130(5):1677–1686, Jun 1963.

[28] P. Bruno and J. P. Renard. Magnetic surface anisotropy of transition metal ultrathin films. Applied Physics A: Materials Science & Processing, 49:499–506, 1989.

[29] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience, 10:186 EP –, 02 2009.

[30] Gregory C. Burgess, Sridhar Kandala, Dan Nolan, Timothy O. Laumann, Jonathan D. Power, Babatunde Adeyemo, Michael P. Harms, Steven E. Petersen, and Deanna M. Barch. Evaluation of denoising strategies to address motion-correlated artifacts in resting-state functional magnetic resonance imaging data from the human connectome project. Brain Connectivity, 6(9):669–680, 2018/01/01 2016.

[31] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. The Annals of Mathematical Statistics, 23:493–507, Dec. 1952.

[32] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino, and Belle L Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 153–162. ACM, 2007.

[33] B. G. Christensen, A. Hill, P. G. Kwiat, E. Knill, S. W. Nam, K. Coakley, S. Glancy, L. K. Shalm, and Y. Zhang. Analysis of coincidence-time loopholes in experimental Bell tests. Phys. Rev. A, 92:032130, Sep 2015.

[34] Fan Chung. Laplacians and the cheeger inequality for directed graphs. Annals of Combinatorics, 9(1):1–19, 2005.

[35] Fan R.K. Chung. Spectral Graph Theory. American Mathematical Society, 1997.

[36] S. Chung, A. Eklund, E. Iacocca, S.M. Mohseni, S.R. Sani, L. Bookman, M. A. Hoefer, R.K. Dumas, and J. Akerman. Magnetic droplet nucleation. Nature Communications, 2015.

[37] Amin Coja-Oghlan. On the laplacian eigenvalues of $G(n, p)$. Combinatorics, Probability and Computing, 16(06):923–946, 2007.

[38] Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, Andr?s Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, Chaogan Yan, and Pierre Bellec. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Frontiers in Neuroinformatics, 41, 2013.

[39] M. d'Aquino, C. Serpico, G. Coppola, I. D. Mayergoyz, and G. Bertotti. Midpoint numerical technique for stochastic landau-lifshitz-gilbert dynamics. J. Appl. Phys., 99(8), 2006.

[40] A Di Martino, C-G Yan, Q Li, E Denio, F X Castellanos, K Alaerts, J S Anderson, M Assaf, S Y Bookheimer, M Dapretto, B Deen, S Delmonte, I Dinstein, B Ertl-Wagner, D A Fair, L Gallagher, D P Kennedy, C L Keown, C Keysers, J E Lainhart, C Lord, B Luna, V Menon, N J Minshew, C S Monk, S Mueller, R-A MÃ¼ller, M B Nebel, J T Nigg, K O'Hearn, K A Pelphrey, S J Peltier, J D Rudie, S Sunaert, M Thioux, J M Tyszka, L Q Uddin, J S Verhoeven, N Wenderoth, J L Wiggins, S H Mostofsky, and M P Milham. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry, 19:659 EP –, 06 2013.

[41] L.B. Holder D.J. Cook, editor. Mining Graph Data. Wiley, 2006.

[42] Claire Donnat and Susan Holmes. Tracking network dynamics: a review of distances and similarity metrics. arXiv preprint, 01 2018.

[43] Tommaso Dorigo. Extraordinary claims: the 0.000029% solution. In EPJ Web of Conferences, volume 95, page 02003. EDP Sciences, 2015.

[44] P.G. Doyle and J.L. Snell. Random walks and electric networks. AMC, 10:12, 1984.

[45] Randy K. Dumas, E. Iacocca, S. Bonetti, S. R. Sani, S. M. Mohseni, A. Eklund, J. Persson, O. Heinonen, and Johan Akerman. Spin-wave-mode coexistence on the nanoscale: A consequence of the oersted-field-induced asymmetric energy landscape. Phys. Rev. Lett., 110:257202, Jun 2013.

[46] R.K. Dumas, S.R. Sani, S.M. Mohseni, E. Iacocca, Y. Pogoryelov, P.K. Muduli, S. Chung, P. Dürrenfeld, and J. Akerman. Recent advances in nanocontact spin-torque oscillators. Magnetics, IEEE Transactions on, 50(6):1–7, June 2014.

[47] Anders Eklund, Stefano Bonetti, Sohrab R. Sani, S. Majid Mohseni, Johan Persson, Sunjae Chung, S. Amir Hossein Banuazizi, Ezio Iacocca, Mikael Östling, Johan Akerman, and B. Gunnar Malm. Dependence of the colored frequency noise in spin torque oscillators on current and magnetic field. Applied Physics Letters, 104(9):–, 2014.

[48] W. Ellens, F.M. Spieksma, P. Van Mieghem, A. Jamakovic, and R.E. Kooij. Effective graph resistance. Linear Algebra and its Applications, 435(10):2491 – 2506, 2011.

[49] P Erős and A Rényi. On random graphs i. Publicationes Mathematicae, 6:290–297, 1959.

[50] Illés J. Farkas. Spectra of "real-world" graphs: Beyond the semicircle law. Physical Review E, 64(2), 2001.

[51] Abraham Flaxman, Alan Frieze, and Trevor Fenner. High Degree Vertices and Eigenvalues in the Preferential Attachment Graph, pages 264–274. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[52] Alex Fornito, Andrew Zalesky, Christos Pantelis, and Edward T. Bullmore. Schizophrenia, neuroimaging and connectomics. NeuroImage, 62(4):2296–2314, 2012.

[53] Joel Friedman and Jean-Pierre Tillich. Wave equations for graphs and the edge-based laplacian. Pacific Journal of Mathematics, 216(2):229–266, 2004.

[54] Marco Genovese. Research on hidden variable theories: A review of recent progresses. Physics Reports, 413(6):319–396, 2005.

[55] A. Ghosh, S. Boyd, and A. Saberi. Minimizing effective resistance of a graph. SIAM Rev., 50(1):37–66, February 2008.

[56] Frédéric Gilbert, Paolo Simonetto, Faraz Zaidi, Fabien Jourdan, and Romain Bourqui. Communities and hierarchical structures in dynamic social networks: analysis and visualization. Social Network Analysis and Mining, 1(2):83–95, 2011.

[57] M. Giustina et al. Significant-loophole-free test of Bell's theorem with entangled photons. Phys. Rev. Lett., 115:250401, Dec 2015.

[58] Wayne Goddard and Ortrud R. Oellermann. Distance in Graphs, pages 49–72. Birkhäuser Boston, Boston, 2011.

[59] Jonathan S Golan. The linear algebra a beginning graduate student ought to know. Springer Science & Business Media, 2012.

[60] Leonardo L Gollo and Michael Breakspear. The frustrated brain: from dynamics on motifs to communities and networks. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1653):20130532, 2014.

[61] Willem H. Haemers and Edward Spence. Enumeration of cospectral graphs. European Journal of Combinatorics, 25(2):199 – 211, 2004.

[62] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference (SciPy2008), pages 11–15, Pasadena, CA USA, August 2008.

[63] N. Halko, P. Martinsson, and J. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review, 53(2):217–288, 2018/02/28 2011.

[64] T. H. Haveliwala. Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. IEEE Transactions on Knowledge and Data Engineering, 15(4):784–796, 2003.

[65] B. Hensen et al. Loophole-free Bell inequality violation using electron spins separated by 1.3 km. Nature, 526:682, 2015.

[66] M. A. Hoefer, T. J. Silva, and Mark W. Keller. Theory for a dissipative droplet soliton excited by a spin torque nanocontact. Phys. Rev. B, 82:054432, Aug 2010.

[67] M. A. Hoefer, M. Sommacal, and T. J. Silva. Propagation and control of nanoscale magnetic-droplet solitons. Phys. Rev. B, 85(21):214433, 2012.

[68] M.A. Hoefer and M. Sommacal. Propagating two-dimensional magnetic droplets. Physica D: Nonlinear Phenomena, 241(9):890 – 901, 2012.

[69] W. Hoeffding. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 58:13, 1963.

[70] N Hoque, Monowar H Bhuyan, Ram Charan Baishya, DK Bhattacharyya, and Jugal K Kalita. Network attacks: Taxonomy, tools and systems. Journal of Network and Computer Applications, 40:307–324, 2014.

[71] Jocelyn V Hull, Zachary J Jacokes, Carinna M Torgerson, Andrei Irimia, and John Darrell Van Horn. Resting-state functional connectivity in autism spectrum disorders: A review. Frontiers in Psychiatry, 7:205, 2016.

[72] Marc-Thorsten Hütt, Marcus Kaiser, and Claus C Hilgetag. Perspective: network-guided pattern formation of neural dynamics. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1653):20130522, 2014.

[73] Ezio Iacocca, Randy K. Dumas, Lake Bookman, Majid Mohseni, Sunjae Chung, Mark A. Hoefer, and Johan Akerman. Confined dissipative droplet solitons in spin-valve nanowires with perpendicular magnetic anisotropy. Phys. Rev. Lett., 112:047201, Jan 2014.

[74] Tsuyoshi Ide and Hisashi Kashima. Eigenspace-based anomaly detection in computer systems. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 440–449. ACM, 2004.

[75] B.A. Ivanov and V.A. Stephanovich. Two-dimensional soliton dynamics in ferromagnets. Physics Letters A, 141(1):89 – 94, 1989.

[76] Márton Karsai, Nicola Perra, and Alessandro Vespignani. Time varying networks and the weakness of strong ties. Scientific reports, 4, 2014.

[77] William L. Kath and Noel F. Smyth. Soliton evolution and radiation loss for the nonlinear schrödinger equation. Phys. Rev. E, 51:1484–1492, Feb 1995.

[78] Alistair King, Bradley Huffaker, Alberto Dainotti, et al. A coordinated view of the temporal evolution of large-scale internet events. Computing, 96(1):53–65, 2014.

[79] D.J. Klein and M. Randić. Resistance distance. Journal of Mathematical Chemistry, 12(1):81–95, 1993.

[80] P.E. Kloeden and E Platen. Numerical Solution of Stochastic Differential Equations. Springer-Verlag Berlin Heidelberg, 1992.

[81] A.M. Kosevich, B.A. Ivanov, and A.S. Kovalev. Magnetic solitons. Physics Reports, 194(34):117 – 238, 1990.

[82] Danai Koutra. Exploring and Making Sense of Large Graphs. PhD thesis, Carnegie Mellon University, 2015. CMU-CS-15-126.

[83] Danai Koutra, Tai-You Ke, U. Kang, Duen Horng (Polo) Chau, Hsing-Kuo Kenneth Pao, and Christos Faloutsos. Unifying guilt-by-association approaches: Theorems and fast algorithms. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, Machine Learning and Knowledge Discovery in Databases, pages 245–260, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[84] Danai Koutra, Neil Shah, Joshua T Vogelstein, Brian Gallagher, and Christos Faloutsos. Delta Con: Principled massive-graph similarity function with attribution. ACM Transactions on Knowledge Discovery from Data (TKDD), 10(3):28, 2016.

[85] Danai Koutra, Joshua T. Vogelstein, and Christos Faloutsos. Deltacon: A principled massive-graph similarity function. ACM Transactions on Knowledge Discovery from Data,, V(N), May 2014.

[86] Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, and Jari Saramäki. Temporal motifs. In Temporal Networks, pages 119–133. Springer, 2013.

[87] S. Kullback and R. A. Leibler. On information and sufficiency. Ann. Math. Statist., 22:79, 1951.

[88] Timothy La Fond, Jennifer Neville, and Brian Gallagher. Anomaly detection in dynamic networks of varying size. arXiv preprint arXiv:1411.3749, 2014.

[89] J.-A. Larsson. Loopholes in Bell inequality tests of local realism. J. Phys. A, 47:424003, 2014. arXiv:1407.0363.

[90] James R. Lee, Shayan Oveis Gharan, and Luca Trevisan. Multiway spectral partitioning and higher-order cheeger inequalities. J. ACM, 61(6):37:1–37:30, December 2014.

[91] J.R. Lee, , S.O. Gharan, and L. Trevisan. Multiway spectral partitioning and higher-order Cheeger inequalities. Journal of the ACM, 61(6):37, 2014.

[92] S. Lendínez, N. Statuto, D. Backes, A.D. Kent, and F. Macià. Observation of droplet soliton drift resonances in a spin-transfer-torque nanocontact to a ferromagnetic thin film. arXiv:1507:08218, 2015.

[93] Sergi Lendínez, Jinting Hang, Saül Vélez, Joan Manel Hernández, Dirk Backes, Andrew D. Kent, and Ferran Macià. Effect of temperature on magnetic solitons induced by spin-transfer torque. Phys. Rev. Applied, 7:054027, May 2017.

[94] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. Markov chains and mixing times. American Mathematical Soc., 2009.

[95] Dean Lusher, Johan Kuskinen, and Garry Robbins, editors. Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications. Structural Analysis in the Social Sciences. Cambridge University Press, 2012.

[96] R. Lyons and Y. Peres. Probability on trees and networks. In preparation. Available at `http://http://mypage.iu.edu/~rdlyons/`http://mypage.iu.edu/ rdlyons, 2005.

[97] Christoph Maas. Computing and interpreting the adjacency spectrum of traffic networks. Journal of Computational and Applied Mathematics, 12-13(Supplement C):459 – 466, 1985.

[98] F. Macià, D. Backes, and A. Kent. Stable magnetic droplet solitons in spin transfer nanocontacts. Nature Nano., 10:1038, 2014.

[99] M. Madami, E. Iacocca, S. Sani, G. Gubbiotti, S. Tacchi, R. K. Dumas, J. Akerman, and G. Carlotti. Propagating spin waves excited by spin-transfer torque: A combined electrical and optical study. Phys. Rev. B, 92:024403, Jul 2015.

[100] M. D. Maiden, L. D. Bookman, and M. A. Hoefer. Attraction, merger, reflection, and annihilation in magnetic droplet soliton scattering. Phys. Rev. B, 89:180409, 2014.

[101] Patrick L. McDermott and Christopher K. Wikle. Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data. arXiv preprint, 11 2017.

[102] Brendan D. McKay. On Littlewood's estimate for the binomial distribution. Adv. Appl. Prob., 21(2):475–478, 1989.

[103] V. Michel, C. Damon, and B. Thirion. Mutual information-based feature selection enhances fmri brain activity classification. In 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pages 592–595, 2008.

[104] Anish Mitra, Abraham Z. Snyder, John N. Constantino, and Marcus E. Raichle. The lag structure of intrinsic activity is focally altered in high functioning adults with autism. Cerebral Cortex, 27(2):1083–1093, 02 2017.

[105] S. M. Mohseni, S. R. Sani, J. Persson, T. N. Anh Nguyen, S. Chung, Ye. Pogoryelov, P. K. Muduli, E. Iacocca, A. Eklund, R. K. Dumas, S. Bonetti, A. Deac, M. A. Hoefer, and J. Akerman. Spin torque generated magnetic droplet solitons. Science, 339(6125):1295–1298, 2013.

[106] S.M. Mohseni, S.R. Sani, R.K. Dumas, J. Persson, T.N. Anh Nguyen, S. Chung, Ye. Pogoryelov, P.K. Muduli, E. Iacocca, A. Eklund, and J. Akerman. Magnetic droplet solitons in orthogonal nano-contact spin torque oscillators. Physica B: Condensed Matter, 435(0):84 – 87, 2014.

[107] Nathan D Monnig and François G Meyer. The resistance perturbation distance: A metric for the analysis of dynamic networks. Discrete and Applied Mathematics, page to appear., 2017.

[108] Nathan D. Monnig and François G. Meyer. The resistance perturbation distance: A metric for the analysis of dynamic networks. CoRR, abs/1605.01091, 2016.

[109] E. F. Moore. The shortest path through a maze. Proceedings of an International Symposium on the Theory of Switching, pages 285–292, 1959.

[110] R. O. Moore, G. Biondini, and W. L. Kath. A method to compute statistics of large, noise-induced perturbations of nonlinear schrodinger solitons. SIAM Review, 50(3):523–549, 2008.

[111] S. V. Nagaev and V. I. Chebotarev. On the bound of proximity of the binomial distribution to the normal one. Doklady Mathematics, 436(1):26–18, 2011.

[112] N. Nagaosa and Y. Tokura. Topological properties and dynamics of magnetic skyrmions. Nature Nano., 8:898, 2013.

[113] Mark Newman. Networks: an introduction. Oxford university press, 2010.

[114] Sean O'Rourke. Gaussian fluctuations of eigenvalues in wigner random matrices. Journal of Statistical Physics, 138(6):1045–1066, 2010.

[115] Christos H. Papadimitriou. Computational complexity. In Encyclopedia of Computer Science, pages 260–265. John Wiley and Sons Ltd., Chichester, UK, 2003.

[116] Jagdish K Patel and Campbell B Read. Handbook of the normal distribution, volume 150. CRC Press, 1996.

[117] S. Petit-Watelot, R. M. Otxoa, and M. Manfrini. Electrical properties of magnetic nanocontact devices computed using finite-element simulations. Applied Physics Letters, 100(8):083507, 2012.

[118] Mark Plitt, Kelly Anne Barnes, and Alex Martin. Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. NeuroImage: Clinical, 7:359–366, 2015.

[119] C. Ragusa, M. d'Aquino, C. Serpico, B. Xie, M. Repetto, G. Bertotti, and D. Ansalone. Full micromagnetic numerical simulations of thermal fluctuations. IEEE Transactions on Magnetics, 45(10):3919–3922, Oct 2009.

[120] Gyan Ranjan, Zhi-Li Zhang, and Daniel Boley. Incremental computation of pseudo-inverse of laplacian. In Proceedings of the 8th International Conference on Combinatorial Optimization and Applications, pages 729–749, 2014.

[121] Stephen Ranshous, Shitian Shen, Danai Koutra, Steve Harenberg, Christos Faloutsos, and Nagiza F Samatova. Anomaly detection in dynamic networks: a survey. Wiley Interdisciplinary Reviews: Computational Statistics, 7(3):223–247, 2015.

[122] Elizabeth Redcay, Joseph Moran, Penelope Mavros, Helen Tager-Flusberg, John Gabrieli, and Susan Whitfield-Gabrieli. Intrinsic functional network organization in high-functioning adolescents with autism spectrum disorder. Frontiers in Human Neuroscience, 7:573, 2013.

[123] Fabio Roli, Giorgio Giacinto, and Gianni Vernazza. Methods for Designing Multiple Classifier Systems, pages 78–87. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[124] N. Romming, C. Hanneken, M. Menzel, J. E. Bickel, B. Wolter, K. von Bergmann, A. Kubetzka, and R. Wiesendanger. Writing and deleting single magnetic skyrmions. Science, 341(6146):636–639, 2013.

[125] Wenjamin Rosenfeld, Daniel Burchardt, Robert Garthoff, Kai Redeker, Norbert Ortegel, Markus Rau, and Harald Weinfurter. Event-ready Bell-test using entangled atoms simultaneously closing detection and locality loopholes. Phys. Rev. Lett., 119:010402, 2017.

[126] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: Uses and interpretations. NeuroImage, 52(3):1059 – 1069, 2010.

[127] W. Rudin. Functional Analysis. International series in pure and applied mathematics. McGraw-Hill, 1991.

[128] J. Sampaio, V. Cros, S. Rohart, A. Thiaville, and A. Fert. Nucleation, stability and current-induced motion of isolated magnetic skyrmions in nanostructures. Nature Nano., 8:839–844, 2013.

[129] A. J. Schwenk. Almost all trees are cospectral. New Directions in the Theory of Graphs, 1973.

[130] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. Statist. Sci., 26:84–101, 2011.

[131] L. K. Shalm et al. Strong loophole-free test of local realism. Phys. Rev. Lett., 115:250402, Dec 2015.

[132] Lynden K. Shalm, Evan Meyer-Scott, Bradley G. Christensen, Peter Bierhorst, Michael A. Wayne, Martin J. Stevens, Thomas Gerrits, Scott Glancy, Deny R. Hamel, Michael S. Allman, Kevin J. Coakley, Shellee D. Dyer, Carson Hodge, Adriana E. Lita, Varun B. Verma, Camilla Lambrocco, Edward Tortorici, Alan L. Migdall, Yanbao Zhang, Daniel R. Kumor, William H. Farr, Francesco Marsili, Matthew D. Shaw, Jeffrey A. Stern, Carlos Abellán, Waldimar Amaya, Valerio Pruneri, Thomas Jennewein, Morgan W. Mitchell, Paul G. Kwiat, Joshua C. Bienfang, Richard P. Mirin, Emanuel Knill, and Sae Woo Nam. Strong loophole-free test of local realism. Phys. Rev. Lett., 115:250402, Dec 2015.

[133] Jun Shao. Mathematical Statistics. Springer New York, 2nd edition, 2003.

[134] T.J. Silva and M.W. Keller. Theory of thermally induced phase noise in spin torque oscillators for a high-symmetry case. Magnetics, IEEE Transactions on, 46(9):3555 –3573, sept. 2010.

[135] J. C. Slonczewski. Current-driven excitation of magnetic multilayers. J. Magn. Magn. Mater., 159(1-2):L1 – L7, 1996.

[136] D. A. Spielman and S.-H. Teng. Spectral sparsification of graphs. SIAM J. Comput., 40(4):981–1025, 2011.

[137] Cornelis J Stam. Modern network science of neurological disorders. Nature Reviews Neuroscience, 15(10):683–695, 2014.

[138] Juliette Stehle, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, et al. High-resolution measurements of face-to-face contact patterns in a primary school. PloS one, 6(8):e23176, 2011.

[139] E. Stokstad. New hints into the biological basis of autism. Science, 294(5540):34–7, Oct. 2001.

[140] Vigneshwaran Subbaraju, Mahanand Belathur Suresh, Suresh Sundaram, and Sundararajan Narasimhan. Identifying differences in brain activities and an accurate detection of autism spectrum disorder using resting state functional-magnetic resonance imaging : A spatial filtering approach. Medical Image Analysis, 35:375–389, 2017.

[141] Yizhou Sun, Jie Tang, Jiawei Han, Manish Gupta, and Bo Zhao. Community evolution detection in dynamic heterogeneous information networks. In Proceedings of the Eighth Workshop on Mining and Learning with Graphs, pages 137–146. ACM, 2010.

[142] Kaustubh Supekar, Vinod Menon, Daniel Rubin, Mark Musen, and Michael D. Greicius. Network analysis of intrinsic functional brain connectivity in alzheimer's disease. PLOS Computational Biology, 4(6):e1000100–, 06 2008.

[143] Yi-Yuan Tang, Mary K. Rothbart, and Michael I. Posner. Neural correlates of establishing, maintaining, and switching brain states. Trends in Cognitive Sciences, 16(6):330–337, 2012.

[144] Riitta Toivonen, Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, and Kimmo Kaski. A model for social networks. Physica A: Statistical Mechanics and its Applications, 371(2):851–860, 2006.

[145] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. NeuroImage, 15(1):273–289, 2002.

[146] M P van den Heuvel, O Sporns, G Collin, and et al. Abnormal rich club organization and functional brain dynamics in schizophrenia. JAMA Psychiatry, 70(8):783–792, 2013.

[147] Arne Vansteenkiste, Jonathan Leliaert, Mykola Dvornik, Mathias Helsen, Felipe Garcia-Sanchez, and Bartel Van Waeyenberge. The design and verification of mumax3. AIP Advances, 4(10):107133, 2014.

[148] J. Ville. Etude Critique de la Notion de Collectif. Gauthier-Villars, Paris, 1939.

[149] Christoph Vogler, Florian Bruckner, Bernhard Bergmair, Thomas Huber, Dieter Suess, and Christoph Dellago. Simulating rare switching events of magnetic nanostructures with forward flux sampling. Phys. Rev. B, 88:134409, 2013.

[150] Ulrike von Luxburg. A tutorial on spectral clustering. Statistics and Computing, 17(4), 11 2007.

[151] Ulrike Von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. Journal of Machine Learning Research, 15(1):1751–1798, 2014.

[152] Ulrike von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. Journal of Machine Learning Research, 15:1751–1798, 2014.

[153] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. Nature, 393(6684):440–442, June 1998.

[154] E. P. Wigner. On the Distribution of the Roots of Certain Symmetric Matrices. The Annals of Mathematics, 67(2):325–327, March 1958.

[155] P. Wills, E. Iacocca, and M. A. Hoefer. Deterministic drift instability and stochastic thermal perturbations of magnetic dissipative droplet solitons. Phys. Rev. B, 93:144408, Apr 2016.

[156] Peter Wills, Emanuel Knill, Kevin Coakley, and Yanbao Zhang. Performance of test supermartingale confidence intervals for the success probability of bernoulli trials. arXiv preprint, 09 2017.

[157] Peter Wills and François G Meyer. Detecting topological changes in dynamic community networks. arXiv preprint arXiv:1957950, 2017.

[158] Peter Wills and François G. Meyer. Detecting topological changes in dynamic community networks. CoRR, abs/1707.07362, 2017.

[159] R.C. Wilson and P. Zhu. A study of graph spectra for comparing graphs and trees. Pattern Recognition, 41(9):2833–2841, 2008.

[160] G. U. Yule. A mathematical theory of evolution, based on the conclusion of dr. j. c. willis, f.r.s. Philisophical Transactions of the Royal Society B., 213:402–410, 1925.

[161] Xiao Zhang, Raj Rao Nadakuditi, and M. E. J. Newman. Spectra of random graphs with community structure and arbitrary degrees. Physical Review E, 89(4):042816–, 04 2014.

[162] Yanbao Zhang, Scott Glancy, and Emanuel Knill. Asymptotically optimal data analysis for rejecting local realism. Phys. Rev. A, 84:062118, Dec 2011.

[163] Yanbao Zhang, Scott Glancy, and Emanuel Knill. Efficient quantification of experimental evidence against local realism. Phys. Rev. A, 88:052119, Nov 2013.

[164] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning from labeled and unlabeled data on a directed graph. In Proceedings of the 22nd International Conference on Machine Learning, ICML '05, pages 1036–1043, New York, NY, USA, 2005. ACM.

[165] Y. Zhou, E. Iacocca, A. Awad, R.K. Dumas, H.B. Zhang, H. B. Braun, and J. Akerman. Dynamically stabilized magnetic skyrmions. Nature Comm., 6(8193), 2015.