Novel Statistical methods and applications in Quantum Computing, Biomedicine, and Networks for Big Data

by

Ye Wang

(Under the Direction of Ping Ma)

Abstract

With the rapid development of science and technology, large and complex data have been generated in many areas, such as social science, neuroscience, and biomedicine. The extraordinary amount of data revolutionize our conventional decision-making system. This new phenomenon poses significant challenges on current statistical research. Therefore, my primary research goals are to develop new theoretically justifiable and computationally efficient methods for tackling big data applications from a computational and modeling perspective. To achieve my goals, a novel non-oracular quantum adaptive search (QAS) method for the best subset selection problem is proposed as the first topic. QAS performs almost identically to the naive best subset selection method but reduces its computational complexity from $O(D)$ to $O(\sqrt{D}log_2 D)$, where $D = 2^p$ is the total number of subset over p covariates. The second topic focuses on a social network application. Drawing on the concepts of community and brokerage from network analysis, we argue that the network of nongovernmental organizations (NGOs) may reinforce power disparities and inequalities at the very same time that it improves access to global governance and provides social power. Finally, the third topic is about biomedicine. The novel statistical analysis is applied to the clinical research on Obstructive Sleep Apnea (OSA). We found that several soluble cytokine receptors are associated with OSA. These findings may facilitate developing new treatment/therapy for patients with OSA.

Index words: Best subset selection, NP-hard problem, non-oracular quantum search, quantum linear prediction, community detection, role analysis, obstructive sleep apnea (OSA)

Novel Statistical methods and applications in Quantum Computing, Biomedicine, and Networks for Big Data

by

Ye Wang

B.E., Guangdong University of Technology, China, 2006
M.S., University of Georgia, United States of America, 2015

A Dissertation Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor OF PHILOSOPHY

Athens, Georgia

2021

Novel Statistical methods and applications in Quantum Computing, Biomedicine, and Networks for Big Data

by

Ye Wang

Major Professor:	Ping Ma
Committee:	Abhyuday Mandal
	Amanda Murdie
	Wenxuan Zhong
	Yuan Ke

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2021

Acknowledgments

This work was made possible through the support of many people. I cannot thank enough my advisor Prof. Ping Ma, who has generously provide me with assistance, guidance, and support throughout my Ph.D. study. I have been greatly impressed and influenced by his insights, wisdom and passion in the statistics. I would like to express my deepest gratitude to him. Meanwhile, I also want to express my sincere gratitude to my committee members: Profs. Abhyuday Mandal, Amanda Murdie, Wenxuan Zhong and Yuan Ke for their advice, help, and encouragement on my dissertation research. It is my honor to have them all on my committee.

I want to appreciate my collaborators, co-authors of my journal papers including Prof. Richard Meagher and Prof. Natarajan Kannan. This work would not be possible without their excellent collaboration and gudiance, especially for Prof. Richard Meagher. I really enjoyed and learned a lot in the collaboration with him.

Moreover, I want to express my thanks to all labmates of Big Data Analytics Lab (BDAL) for their supports: Xiaoxiao Sun, Yiwen Liu, Xin Xing, Xinlian Zhang, Wei Xu, Cheng Meng, Jingyi Zhang, Mengrui Zhang, Jinyang Chen, Lexiang Ji, Huimin Cheng, Yongkai Chen and many others whose names are not listed. It was fantastic to have the opportunity to work with these young researchers in the past five years. Meanwhile, I benefited tremendously from intellectual discussions with colleagues from Department of Statistics and other departments at the University of Georgia, including Zhengbo Ma, Xianyan Chen, and Liang-Chin Huang. I would also like to express my genuine gratitude to the faculty members in the Department of Statistics at the University of Georgia because they have truly influenced me through their extraordinary dedication to teaching and molding of young minds.

This dissertation research also received support from many staff members at Department of Statistics, including Wendy Starnes and Melissa Pettigrew. Part of the dissertation research is supported by U.S. National Science Foundation grants DMS-1903226 to PI Wenxuan Zhong and DMS-1925066 to PI Ping Ma.

Table of Contents

List of Figures

List of Tables

Chapter 1

Overview

With the rapid development of science and technology, large and complex data have been generated in many areas, such as genomics, neuroscience, and social science. The extraordinary amount of data land for artificial intelligence and revolutionize our conventional decision making system. This new phenomenon posts great challenges on current statistical research. For example, the ultra-large size of dataset renders the application of many statistical methods computationally infeasible. Developing new theoretically justifiable and computationally efficient methods for tackling big data problems from a computational and modeling perspective is the primary goal for my research. To achieve my goals in this thesis,

1. In Chapter 2, I develop quantum computing algorithm to overcome computing bottleneck for NP-hard problem in big data environment.
2. In Chapter 3, I borrow role identification concepts after community detection to further understand the mechanism of network evolution based on massive network
3. In Chapter 4, I apply statistical method to identify four cytokines associated with autoimmune disease. These findings may facilitate developing new treatment/therapy for patients with Obstructive Sleep Apnea (OSA).

In Chapter 2, I investigate solving the best subset selection problem by quantum computation. Specifically, it proposes a quantum adaptive search (QAS) algorithm to solve the best subset selection problem. Why choosing quantum computation? Classical computers only manipulate ones and zeros through bit operations. Unlike classical computers, quantum computers use qubits(quantum bits) which have a third state called "superposition". The superposition could represent a one or a zero at the same time. For example, a

quantum system with p qubits can be in any superposition of 2^p different states simultaneously, while a classical system with p bits can only be in one state of 2^p different states at a time. With this great advantage, Quantum computers can solve problems that are impossible or would take a classical computer an impractical amount of time to solve, such as NP-hard problem caused by best subset selection. In order to overcome the computational bottleneck caused by best subset selection methods, we investigate the possibility of solving the best subset selection problem on a quantum computing system.

Although quantum mechanism has motivated significant developments of scalable quantum algorithms in many areas such as algebraic number theory, scientific simulations, optimization, and many more, existing quantum search algorithms are not tailored for statistical learning problems like the best subset selection. We [10] propose a novel non-oracular quantum adaptive search (QAS) method for the best subset selection problems. QAS performs almost identical to the naive best subset selection method but reduces its computational complexity from $O(D)$ to $O(\sqrt{D}\log_2 D)$, where $D = 2^p$ is the total number of subsets over p covariates. Unlike existing quantum search algorithms, QAS does not require the oracle information of the true solution state and hence is applicable to various statistical learning problems with random observations. Theoretically, we prove QAS attains any arbitrary success rate $q \in (0.5, 1)$ within $O(\log_2 D)$ iterations. When the underlying regression model is linear, we propose a quantum linear prediction method which is faster than its classical counterpart. We further introduce a hybrid quantum-classical strategy to avoid the capacity bottleneck of existing quantum computing systems and boost the success rate of QAS by majority voting. The effectiveness of this strategy is justified by both theoretical analysis and extensive empirical experiments on quantum and classical computers.

In Chapter 3, I focus on social network analysis. A social network consists of a set of actor(node) and a set of relationships(edges) between them which describe certain patterns of interaction. Classifying actors by roles they are playing in the network can give us a comprehensive view of the network and can help us understand how it is organized, and even to predict how it could behave in a specific event. In this chapter, we have a nongovernmental organizations (NGOs) network, which is built on the dataset of the $3,903$ NGOs connected through 1.3 million ties occurring through meetings and conferences for NGOs put on or coordinated by the United Nations (UN).

We first outline how community detection helps us understand emergent divisions within NGOs network with 1.3 millions edges. Then we use a time-varying stochastic block model to identify the optimal common composition

of communities that maximizes the modularity across years. The concept of modularity was introduced for networking clustering in 2004 by M. E. J. Newman (Newman, 2006). Next, we then turn our focus to brokerage and brokerage roles (coordinator, itinerant broker, gatekeeper/representative, liaison) of each node/organization. These concepts can significantly help us understand how the advocacy network can extend power. Some communities are likely to have more coordinator brokers, connecting organizations mainly within a specific community, while other communities could have more itinerant, gatekeeper/representative, or liaison brokers, connecting organizations outside of their specific community in divergent ways. In this way, brokerage provides power to organizations differently across communities.

In Chapter 4, I turn my attention to clinical data analysis related to Obstructive sleep apnea. Obstructive sleep apnea (OSA) is a sleep-related breathing disorder associated with numerous adverse health effects. It is estimated that 22 million Americans suffer from sleep apnea, with 80 percent of the cases of moderate and severe obstructive sleep apnea undiagnosed. OSA patients may display any one or several symptoms including fragmented sleep, snoring, excessive daytime sleepiness, fatigue, high blood pressure, irritability, depression, memory loss, and loss of concentration. These health problems develop over months and years and include cardiovascular disease, metabolic syndrome, kidney disease, autoimmune diseases, and neurodegenerative disease (ND). In this chapter,

1. Identification: We identified four cytokines (APRIL, CD30, IFN-Alpha-2 and IL-2) associated with autoimmune disease from 37 inflammatory cytokines candidates. We can find that cytokine levels in airways treated patients were similar to the levels in control subjects.
2. Visualization: The techniques for dimension reduction based on high dimensional patient cytokine data, t-SNE and UMAP, identify only two groups in 2-dimensional space. Group 1 represents all the untreated OSA patients. Group 2 represents all the control individuals and 80% of airways treated OSA patients.
3. Therapeutic Treatment: New treatment/therapy may be developed for patient with Obstructive sleep apnea based on those four identified cytokines.

Chapter 2

Novel Statistical Methods in Quantum Computing for Best Subset Selection

2.1 Introduction

Best subset selection has been a classical and fundamental method for linear regressions (e.g. Beale et al., 1967; Fan et al., 2020; Hocking & Leslie, 1967; Shen et al., 2012). When the covariates contain redundant information for the response, selecting a parsimonious best subset of covariates may improve the model interpretability as well as the prediction performance.

The best subset selection for linear regression can be formulated as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \le t, \tag{2.1}$$

where $\boldsymbol{Y}=(y_1,\ldots,y_n)^\intercal\in\mathbb{R}^n$ is a response vector, $\boldsymbol{X}=(x_{i,j})_{1\le i\le n,1\le j\le p}\in\mathbb{R}^{n\times p}$ is a design matrix of covariates and $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^\intercal\in\mathbb{R}^p$ is a vector of unknown regression coefficients. Throughout this chapter, we assume $\boldsymbol{Y}$ and $\boldsymbol{X}$ are centered for the simplicity of presentation. Further, $\|\boldsymbol{\beta}\|_0$ is the ℓ_0 norm of $\boldsymbol{\beta}$ and $0\le t\le \min(n,p)$ is a subset size. In general, solving (2.1) is a non-convex optimization problem. Without a convex relaxation or a stage-wise approximation, the optimization involves a combinatorial search over all subsets of sizes $\{1,\ldots,\min(n,p)\}$ and hence is an NP-hard problem (Natarajan, 1995; Welch, 1982). Therefore, seeking the exact solution of (2.1) is usually computationally intractable when p is moderate or large.

To overcome the computational bottleneck of the best subset selection, many alternative and approximate approaches have been developed and carefully studied. Stepwise regression methods (e.g. Draper & Smith, 1966; Efroymson, 1960) sequentially add or eliminate covariates according to their marginal contributions to the response condition on the existing model. Stepwise regression methods speed up the best subset selection as they only optimize (2.1) over a subset of all candidates, and hence their solutions may not coincide with the solution of (2.1). Besides, stepwise regression methods suffer from the instability issue in high dimensional regime (Breiman, 1996). Hard thresholding pursuit (e.g. Blumensath et al., 2007; Fan et al., 2020; X.-T. Yuan et al., 2017) proposes an iterative greedy selection procedure to find sparse solutions for underdetermined linear systems. Regularized regression methods suggest to relax the nonconvex ℓ_0 norm in (2.1) by various convex or nonconcave alternatives. A celebrated member in the house is LASSO (Tibshirani, 1996) which solves an ℓ_1 regularized problem instead. Elegant statistical properties and promising numerical performance soon made regularized regression a popular research topic in statistics, machine learning and other data science related areas. We refer to basis pursuit (Chen & Donoho, 1994), SCAD (Fan & Li, 2001), elastic net (Zou & Hastie, 2005), adaptive LASSO (Zou, 2006), Dantzig selector (Candes & Tao, 2007), nonnegative garrotte (M. Yuan & Lin, 2007), and MCP (C.-H. Zhang, 2010), among many others. Though regularized regression methods can perform equivalent to or even better than the solution of (2.1) in certain scenarios (e.g. Fan et al., 2014; Hastie et al., 2020; P. Zhao & Yu, 2006), they are not universal replacements for the best subset selection method. Nevertheless, solving (2.1) efficiently remains a statistically attractive and computationally challenging problem.

Recently, (Bertsimas et al., 2016) has reviewed the best subset selection problem from a modern optimization point of view. Utilizing the algorithmic advances in mixed integer optimization, (Bertsimas et al., 2016) proposed a highly optimized solver of (2.1) which is scalable to relatively large n and p. Later, (Hazimeh & Mazumder, 2018) considered a new hierarchy of necessary optimality conditions and propose to solve (2.1) by adding extra convex regularizations. Then, Hazimeh and Mazumder, 2018 developed an efficient algorithm based on coordinate descent and local combinatorial optimization. Such recent optimization developments push the frontier of computation for the best subset selection problems by a big margin. Despite the NP-hardness of the best subset selection problem has not been improved, an up-to-date electronic computer together with a state-of-the-art optimization algorithm can solve (2.1) with hundreds or even thousands of covariates.

In this chapter, we investigate the possibility of solving the best subset selection problem on a quantum computing system. Unlike electronic computers, a quantum computer operates on quantum processing units, or qubits, which can take values 0, 1, or both simultaneously due to the superposition property. The number of complex numbers required to characterize quantum states usually grows exponentially with the size of the system. For example, a quantum system with p qubits can be in any superposition of 2^p orthonormal states simultaneously, while a classical system can only be in one state at a time (Nielsen & Chuang, 2010). Such a paradigm change has motivated significant developments of scalable quantum algorithms in many areas such as algebraic number theory (Hallgren, 2002; Shor, 1999; Van Dam & Shparlinski, 2008), scientific simulations (D. S. Abrams & Lloyd, 1997; Byrnes & Yamamoto, 2006; Hu & Wang, 2020; Kassal et al., 2008; Y. Wang, 2011), optimization (Harrow et al., 2009; Jansen et al., 2007; S. P. Jordan, 2005), and many more.

However, quantum algorithms tackling the best subest selection problem are still lacking. The most closely related algorithms are quantum search algorithms (e.g. Boyer et al., 1998; Grover, 1997; Høyer et al., 2002; Kwiat et al., 2000; Long, 2001). There are significant challenges to apply these algorithms to subset selection problems. On one hand, they are designed to search an item from a non-random database such as a set of fixed numbers or solutions for a Sudoku problem. Thus, a quantum search algorithm may fail when it is applied to a random set that depends on the observed sample. On the other hand, existing quantum search algorithms are, in general, oracular algorithms that depend on an oracle to decide if an item is a solution or not. For example, Grover's algorithm (Grover, 1997) requires an oracle function that can map all solution states to 1 and all non-solution states to 0 with one operation. However, such a piece of oracle information is usually not available in statistics as the solution is usually a function of random observations.

When the oracle is inaccurate, Grover's algorithm can perform as bad as a random guess which will be demonstrated in our numerical experiments.

To overcome the aforementioned limitations, we propose a novel non-oracular quantum method named quantum adaptive search (QAS). QAS starts with an equally weighted superposition of all candidate models and iteratively updates the superposition towards the direction that minimizes the ℓ_2 loss in (2.1). Within each iteration, the new superposition is compared with the old one through a local evaluation function which does not require any oracle information of the true solution. For a best subset selection problem over $D = 2^p$ candidate models, an electrical computing algorithm requires $O(D)$ quires to search the best model. In contrast, within $O(\log_2 D)$ iterations, QAS con-

verges to a superposition that heavily weighs on the solution state and hence output the exact solution of (2.1) with a high probability. The computational complexity of QAS is upper bounded by the order $O(\sqrt{D}\log_2 D)$ which is only a $\log_2 D$ factor larger than the theoretical lower bound for oracular quantum search algorithms (Bennett et al., 1997). Though the NP-hardness has not been fully conquered, QAS has made a steady step to downscale the computational complexity of the best subset selection problem. When the underlying regression model is linear, we propose a quantum linear prediction method which is faster than its classical counterpart. We further introduce a hybrid quantum-classical strategy to avoid the capacity bottleneck of existing quantum computing systems and boost the success probability of QAS by majority voting.

The rest of the chapter is organized as follows. We review the state-of-the-art quantum search algorithm and its limitation in Section 2.2. In Section 2.3, we propose a novel non-oracular quantum search algorithm named quantum adaptive search (QAS), and discuss its intuition and theoretical properties. In Section 2.4 , we introduce a quantum linear prediction algorithm and suggest a quantum-classical hybrid strategy to improve the stability of quantum best subset selection. In Sections 2.5 and 2.6, we evaluate the empirical performance of the proposed method via extensive experiments on quantum and classical computers. Section 2.10 concludes the chapter with a few remarks. Due to the limitation of space, the proofs of main theoretical results, additional simulation results, and a real data application are relegated to a supplemental material.

2.2 Review of quantum search algorithm

2.2.1 Notations and definitions

To facilitate the discussion in the chapter, we review some essential notations and definitions used in quantum computing which may be alien to the audience of statistics. We refer to (Nielsen & Chuang, 2010) and (Schuld & Petruccione, 2018) for more detailed tutorials of quantum computing. We start by introducing the linear algebra notations used in quantum mechanics, which were invented by Paul Dirac. In Dirac's notation, a column vector $\mathbf{a}$ is denoted by $|a\rangle$ which reads as $\mathbf{a}$ in a 'ket'. Typical vector space of interest in quantum computing is a Hilbert space $\mathcal{H}$ of dimension $D = 2^p$ with a positive integer p. Further, the dual space of $\mathcal{H}$ is defined as follows.

Definition 2.2.1. *For a Hilbert space $\mathcal{H}$, the dual Hilbert space $\mathcal{H}^*$ is defined as the set of linear maps $\mathcal{H} \to \mathbb{C}$, where $\mathbb{C}$ is the complex space.*

For a vector $|a\rangle \in \mathcal{H}$, we denote its dual vector as $\langle a|$ (reads as $\mathbf{a}$ in a 'bra'), which is an element of $\mathcal{H}^*$. Besides, $\mathcal{H}$ and $\mathcal{H}^*$ together naturally induce an inner product $\langle a|b\rangle = \langle \mathbf{a}, \mathbf{b}\rangle$, which is also known as a 'bra-ket'. We say $|a\rangle$ is a unit vector if $\langle a|a\rangle = 1$. The orthonormal basis of $\mathcal{H}$ can be defined as below.

Definition 2.2.2. *For a Hilbert space $\mathcal{H}$ of dimension $D = 2^p$, a set of D vectors $\mathcal{B} = \{|b_0\rangle, |b_1\rangle, \ldots, |b_{D-1}\rangle\}$ is called an orthonormal basis of $\mathcal{H}$ if*

$$\langle b_i|b_j\rangle = \delta_{i,j}, \quad \forall\, b_i, b_j \in \mathcal{B},$$

where $\delta_{i,j} = 1$ when $i = j$ and $\delta_{i,j} = 0$ otherwise.

The framework of quantum computing resides in a state-space postulate which describes the state of a system by a unit vector in a Hilbert space. Recall that, a classical bit in an electronic computer can only store one state at a time over two possibilities (i.e. 0 or 1). In contrast, a state of a quantum bit (or qubit) can be described by a complex unit vector $|\psi\rangle$ in a two-dimensional Hilbert space. It is usually convenient to represent this quantum state as an ancilla qubit

$$|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle,$$

which corresponds to a point on a three-dimensional sphere, known as the Bloch sphere. A visualized comparison between a classical bit and a qubit is given in Figure 2.1.

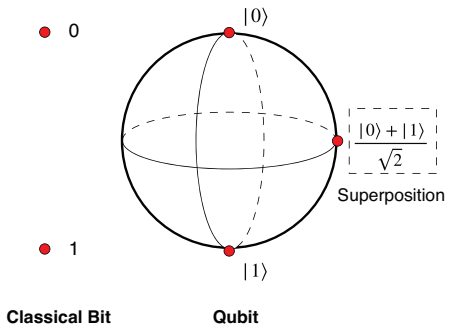

Figure 2.1: Classical bit v.s. qubit

Similarly, a quantum computer of p qubits can represent a state of a system by a unit vector $|\psi\rangle$ in a $D = 2^p$ dimensional Hilbert space $\mathcal{H}$. Let $\mathcal{B} = \{|b_i\rangle\}_{i=0}^{D-1}$ be an orthonormal basis of $\mathcal{H}$. Every unit state $|\psi\rangle \in \mathcal{H}$ can be represented as

$$|\psi\rangle = \sum_{i=0}^{D-1} \phi_i |b_i\rangle, \tag{2.2}$$

where $\phi_0, \ldots, \phi_{D-1}$ is a set of coefficients with $\phi_i = \langle b_i|\psi\rangle$ and $\sum_{i=0}^{D-1} \phi_i^2 = 1$.

Another salient feature of quantum computing is the measurement of a quantum system yields a probabilistic outcome rather than a deterministic one. Suppose the superposition $|\psi\rangle$ in (2.2) is measured, it collapses to a random state in $\mathcal{B}$. In addition, the probability we observe $|b_i\rangle$ is $|\phi_i|^2$ for $i = 0, \ldots, D-1$.

2.2.2 Grover's algorithm

We shall now review the state-of-the-art quantum algorithm to search a particular state in a system of $D = 2^p$ states. For example, searching the smallest real number from a set of D real numbers, or searching a word from a vocabulary list of D words. For the simplicity of discussion, we assume the solution to this searching problem is unique.

In the framework of quantum computing, each state of the system can be modeled by a vector in an orthonormal basis $\mathcal{D} = \{|i\rangle\})_{i=0}^{D-1}$ of a D dimensional Hilbert space $\mathcal{H}$. The solution state is denoted by $|k\rangle$ for some $k \in \{0, \ldots, D-1\}$. (Grover, 1997) proposed a quantum search algorithm with a success probability of at least 50%. The computational complexity of Grover's algorithm is of the order $O(\sqrt{D})$. (Bennett et al., 1997) showed that this rate is optimal, up to a constant, among all possible quantum search algorithms. In contrast, any classical search algorithm needs to query the system for at least $0.5D$ times to solve the searching problem with a 50% or higher success probability. As a result, most classical search algorithms have the computational complexity of the order $O(D)$.

Grover's algorithm assumes that there exists an oracle evaluation function $S(\cdot)$, such that $S(|k\rangle) = 1$ and $S(|i\rangle) = 0$ for $i \neq k$. Grover's algorithm is initialized with a superposition as the equally weighted average of quantum states $|i\rangle, i = 0, \ldots, D-1$. To be specific, the initial superposition is defined as

$$|\psi_0\rangle = \frac{1}{\sqrt{D}} \sum_{i=0}^{D-1} |i\rangle \equiv c_0 |k\rangle + d_0 \sum_{i \neq k} |i\rangle ,$$

where $c_0 = d_0 = \frac{1}{\sqrt{D}}$. Then, Grover's algorithm updates c and d in an iterative manner. In the jth iteration, Grover's algorithm applies the following two operations to the current superposition $|\psi_{j-1}\rangle = c_{j-1} |k\rangle + d_{j-1} \sum_{i \neq k} |i\rangle$,

(a) A flip operation F to $|\psi_{j-1}\rangle$, where $F|i\rangle = (1 - 2S(|i\rangle))|i\rangle$. In other words, $F|k\rangle = -|k\rangle$ and $F|i\rangle = |i\rangle$ for $i \neq k$.

(b) A Grover's diffusion operation G to $F|\psi_{j-1}\rangle$, where $G = 2|\psi_0\rangle\langle\psi_0| - \boldsymbol{I}_D$ and $\boldsymbol{I}_D$ is a $D \times D$ identity matrix.

In the rest of this chapter, we call the two operations together, i.e. GF, as Grover's operation

After the jth iteration, the superposition $|\psi_{j-1}\rangle$ is updated to

$$|\psi_j\rangle = GF|\psi_{j-1}\rangle = c_j|k\rangle + d_j \sum_{i\neq k} |i\rangle,$$

where the coefficients c_j and d_j satisfies

$$\begin{cases} c_j = \frac{D-2}{D} c_{j-1} + \frac{2(D-1)}{D} d_{j-1}, \\ d_j = \frac{D-2}{D} d_{j-1} - \frac{2}{D} c_{j-1}. \end{cases}$$

Let θ be the angle that satisfies $\sin^2\theta = \frac{1}{D}$. The coefficients c_j and d_j admit a closed form (Nielsen & Chuang, 2010),

$$\begin{cases} c_j = \sin\big((2j+1)\theta\big), \\ d_j = \frac{1}{\sqrt{D-1}} \cos\big((2j+1)\theta\big). \end{cases}$$

This closed form provides an intuitive geometric interpretation of Grover's algorithm. Let $|\zeta\rangle = \frac{1}{\sqrt{D-1}} \sum_{i\neq k} |i\rangle$ be the average of all non-solution states which is orthogonal to the solution state $|k\rangle$. It is easy to check that the angel between the initial superposition $|\psi_0\rangle$ and $|\zeta\rangle$ is θ. In the first Grover's operation, the step (a) transforms $|\psi_0\rangle$ to $F|\psi_0\rangle$, which is equivalent to reflect $|\psi_0\rangle$ with respect to $|\zeta\rangle$. In the step (b), the Grover's diffusion operation G reflects $F|\psi_0\rangle$ with respect to $|\psi_0\rangle$. As a result, the angle between $GF|\psi_0\rangle$ and $|\zeta\rangle$ is 3θ. The Figure 2.2 below provides a visualization of the two steps in the first Grover's operation.

Similarly, every iteration in Grover's algorithm is equivalent to rotate the superposition $|\psi_j\rangle$ towards the solution state $|k\rangle$ by 2θ. When D is large, i.e., θ is small, we can approximate the angle by $\theta \approx \sin\theta = \frac{1}{\sqrt{D}}$. Then, a natural stopping criterion for Grover's algorithm is to choose the number of iterations τ by $(2\tau+1)/\sqrt{D} = \pi/2$, which yields τ is approximately $\lceil\sqrt{D}\pi/4\rceil$, where $\lceil\cdot\rceil$ is the ceiling function. Grover's algorithm is summarized in Algorithm 1 below. Grover's algorithm can be easily extend to a search problem with multiple solutions by changing the stopping criterion to $\tau(M) = \lceil\sqrt{D/M}\pi/4\rceil$, where M is the number of solutions. We refer to (Boyer et al., 1998) for more detailed discussions.

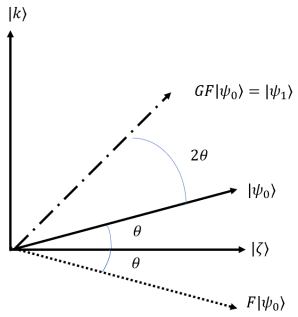

Figure 2.2: Geometrical visualization of the two steps in the first Grover's operation.

Algorithm 1 Grover's algorithm

Input: An orthonormal basis $\mathcal{D}$ of size $D = 2^p$, a binary evaluation function S associated with an oracle state $|k\rangle$, such that $S(|k\rangle) = 1$ and $S(|i\rangle) = 0$ for $i \neq k$. Number of iterations $\tau = \lceil \pi\sqrt{D}/4 \rceil$, where $\lceil \cdot \rceil$ is the ceiling function.

Initialization Prepare an equally weighted superposition on a quantum register of p-qubits $|\psi_0\rangle = \frac{1}{\sqrt{D}} \sum_{i=0}^{D-1}$.

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

Grover's operation: Let $|\psi_j\rangle = GF\,|\psi_{j-1}\rangle$, where $F\,|i\rangle = (1 - 2S(|i\rangle))\,|i\rangle$,

$G = 2\,|\psi_0\rangle\,\langle\psi_0| - \boldsymbol{I}_D$, and $\boldsymbol{I}_D$ is a $D \times D$ identity matrix.

end for

Output: Measure the latest superposition $|\psi_\tau\rangle$ on the quantum register.

2.2.3 Limitation of Grover's algorithm

Grover's algorithm is an oracular quantum algorithm as it depends on an oracle evaluation function that maps the solution state to 1 and all other states to 0. However, such a piece of oracle information is usually not available in statistical learning problems as the states are measured over random samples.

When we have partial or inaccurate oracle information of the solution state, say we may only identify the solution state up to a subset of states, i.e. $|k\rangle \in \mathcal{M} \subset \{|0\rangle, \ldots, |D-1\rangle\}$, the best oracle evaluation function that we can construct is

$$\begin{cases} S(|i\rangle) = 1, & \text{when } i \in \mathcal{M}, \\ S(|i\rangle) = 0, & \text{when } i \in \mathcal{M}^c. \end{cases}$$

Then, each iteration of Grover's algorithm rotates the current superposition towards the hyper-plane spanned by the states in $\mathcal{M}$ instead of the true solution state $|k\rangle$. As a result, Grover's algorithm may fail to converge to the truth and provide a biased estimator. The bias is lower bounded by the difference between $|k\rangle$ and the projection of the initial superposition on the hyper-plane spanned by the states in $\mathcal{M}$.

When we have no oracle information of the solution state at all, the oracle evaluation function can only be constructed by a randomly selected solution state. Then, Grover's algorithm is highly likely to rotate the initial superposition towards a wrong direction and the output of the algorithm can be as bad as a random guess. We empirically demonstrate this phenomenon in Section 2.5.2.

2.3 Best subset selection with quantum adaptive search

2.3.1 Quantum adaptive search

In this subsection, we propose a novel non-oracular quantum search algorithm named quantum adaptive search (QAS) which aims to overcome the aforementioned limitation of Grover's algorithm. Let us consider the best subset selection problem (2.1) and assume $n \geq p$ such that all $D = 2^p$ subsets of p covariates are attainable. When $n < p$, we only need to search over $\sum_{t=0}^{n} \binom{p}{t} < 2^p$ subsets which is a less challenging combinatorial search problem.

To implement the combinatorial search on a quantum computing system, we encode each subset of $\{1, \ldots, p\}$ as a state in an orthonormal basis $\mathcal{D} = \{|0\rangle, \ldots, |D-1\rangle\}$, which consists of $D = 2^p$ elements of a Hilbert space $\mathcal{H}$. Further, we define a state loss function $g(\cdot) : \mathcal{H} \to \mathbf{R}$ as follows

$$g(|i\rangle) := L_n(t; j_1, \cdots, j_t; \widehat{\beta}_{j_1}, \ldots, \widehat{\beta}_{j_t}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{l=1}^{t} \widehat{\beta}_{j_l} x_{i,j_t} \right)^2, \tag{2.3}$$

where the state $|i\rangle \in \mathcal{D}$ is a vector in $\mathcal{H}$ that corresponds to the subset $\{j_1, \cdots, j_t\}$, and $\widehat{\beta}_{j_1}, \ldots, \widehat{\beta}_{j_t}$ are the regression coefficient estimates obtained by minimizing (2.1) with fixed t and $\{j_1, \cdots, j_t\}$. Intuitively, the best subset corresponds to the state that minimizes the state loss function.

Next, we introduce the proposed non-oracular quantum adaptive search method, i.e. QAS, by three key steps as follows.

(1) INITIALIZATION: We choose an initial benchmark state $|w\rangle$ randomly from the set $\mathcal{D} = \{|0\rangle, \ldots, |D-1\rangle\}$. Also, we pre-specify a learning rate $\lambda \in (0, 1)$.
(2) UPDATING: We run Algorithm 1 on set $\mathcal{D}$ by inputting the benchmark state $|w\rangle$ and the number of iterations $\tau = \lceil \pi\lambda^{-m/2}/4 \rceil$, where m is a positive integer. Denote the output of Algorithm 1 as $|w^{new}\rangle$ which is a state in $\mathcal{D}$. Then, we compare $|w^{new}\rangle$ with $|w\rangle$ in terms of the state loss function $g(\cdot)$. If $g(|w^{new}\rangle) < g(|w\rangle)$, we update the current benchmark state $|w\rangle$ with $|w^{new}\rangle$, otherwise we do not update $|w\rangle$.
(3) ITERATION AND OUTPUT: Start with $m = 1$ and repeat the updating step. After each updating step, set $m = m + 1$. The QAS stops when $m > C(\lambda) \ln D$, where $C(\lambda)$ is a positive constant that depends on the learning rate λ. Then, we measure the quantum register with the latest superposition. The output is the observed state in $\mathcal{D}$ and its corresponding subset.

Unlike Grover's algorithm, QAS randomly selects a state in $\mathcal{D}$ as the benchmark state which does not require any oracle information of the solution state. Then, QAS iteratively updates the benchmark state towards the direction that reduces the state loss function. Besides, QAS is data-adaptive in the sense that it starts with a conservative learning step size (e.g. $m = 1$) and gradually increase the learning step size as the benchmark state has been updated towards the truth. We summarize QAS in Algorithm 2 below.

Algorithm 2 Quantum adaptive search

Input: An orthonormal basis $\mathcal{D}$ of size $D = 2^p$, a state loss function $g(\cdot)$ that maps a state in $\mathcal{D}$ to a real number, a learning rate $\lambda \in (0, 1)$.
Initialization Set $m = 1$. Randomly select a state in $\mathcal{D}$ as the initial benchmark state $|w\rangle$. Define a local evaluation function $S(\cdot, |w\rangle, g)$ such that $S(|i\rangle, |w\rangle, g) = 1$ if $g(|i\rangle) \leq g(|w\rangle)$ and $S(|i\rangle, |w\rangle, g) = 0$ if $g(|i\rangle) > g(|w\rangle)$.
repeat
(1) Run Algorithm 1 by inputting $\mathcal{D}$, $S(\cdot, |w\rangle, g)$ and $\tau(m) = \lceil \pi\lambda^{-m/2}/4 \rceil$.
(2) Measure the quantum register and denote the readout by $|w^{new}\rangle$.
(3) If $g(|w^{new}\rangle) < g(|w\rangle)$, set $|w\rangle = |w^{new}\rangle$ and update $S(\cdot, |w\rangle, g)$ accordingly.
(4) $m = m + 1$.
until $m > C(\lambda) \ln D$, where $C(\lambda)$ is a positive constant depends on λ.
Output: The latest benchmark state $|w\rangle$.

Algorithm 2 provides a general non-oracular quantum computing framework for best subset selection problems as the state loss function $g(\cdot)$ can be

tailored for various statistical models, such as nonlinear regression, classification, clustering, and low-rank matrix recovery.

2.3.2 Intuition of quantum adaptive search

In this subsection, we demonstrate the intuition of QAS. Suppose that we implement QAS on an orthonormal basis $\mathcal{D} = \{|i\rangle\}_{i=0}^{D-1}$ with a pre-specified state loss function $g(\cdot) : \mathcal{H} \to \mathbf{R}$ and a learning rate $\lambda \in (0, 1)$. We assume there exists a sole solution state in $\mathcal{D}$ that minimizes $g(\cdot)$. Without loss of generality, we number the states in $\mathcal{D}$ as the ascending rank of their state loss function values, i.e.

$$g(|0\rangle) < g(|1\rangle) \leq g(|2\rangle) \leq \cdots \leq g(|D-1\rangle), \tag{2.4}$$

where $|0\rangle$ is the sole solution state.

If the initialization step in Algorithm 2 luckily selects the sole solution state $|0\rangle$ as the initial benchmark state, QAS will never update the benchmark state and hence reduces to Grover's algorithm with a known oracle state and $\tau \approx \sqrt{D}\pi/4$. Therefore, QAS can recover the true state with a high success probability.

A more interesting discussion would be considering the initial benchmark state $|w\rangle$ does not coincide with the truth, i.e. $|w\rangle \neq |0\rangle$. Given the rank in (2.4), the local evaluation function $S(|i\rangle, |w\rangle, g)$ can be simplified as

$$\begin{cases} S(|i\rangle, |w\rangle, g) = 1, & \text{if } i \leq w, \\ S(|i\rangle, |w\rangle, g) = 0, & \text{if } i > w. \end{cases}$$

In the mth iteration, QAS calls Algorithm 1 by inputting $\mathcal{D}$, $S(\cdot, |w\rangle, g)$ (suppose that $w \neq 0$) and $\tau(m) = \lceil \pi\lambda^{-m/2}/4 \rceil$. Algorithm 1 initializes a equally weighted superposition as

$$|\psi_0\rangle = \frac{1}{\sqrt{D}} \sum_{i=0}^{D-1} |i\rangle \equiv \alpha_0 \sum_{i=0}^{w} |i\rangle + \beta_0 \sum_{j=w+1}^{D-1} |j\rangle, \quad \text{with} \quad \alpha_0 = \beta_0 = \frac{1}{\sqrt{D}}.$$

Then, Algorithm 1 applies $\tau(m)$ Grover's operations to $|\psi_0\rangle$ which updates $|\psi_0\rangle$ to $|\psi_{\tau(m)}\rangle$ with coefficients satisfy

$$\begin{cases} \alpha_{\tau(m)} = \frac{1}{\sqrt{w+1}} \sin\big((2\tau(m)+1)\theta\big), \\ \beta_{\tau(m)} = \frac{1}{\sqrt{D-w-1}} \cos\big((2\tau(m)+1)\theta\big), \end{cases}$$

where the angle θ satisfies $\sin^2\theta = (w+1)/D$. After the mth iteration, Algorithm 1 outputs a random state $|w_{new}\rangle \in \mathcal{D}$ with the following probability mass function

$$P(|w_{new}\rangle = |i\rangle) = \begin{cases} \alpha^2_{\tau(m)}, & \text{if } i \leq w, \\ \beta^2_{\tau(m)}, & \text{if } i > w. \end{cases}$$

Since the learning rate $\lambda \in (0,1)$, we have $\alpha^2_{\tau(m)} > 1/D > \beta^2_{\tau(m)}$ for some positive m. After the mth iteration, QAS amplifies the probability of drawing the states whose state loss function values are smaller or equal to $g(|w\rangle)$. Meanwhile, QAS suppresses the probability of drawing the states whose state loss function values are greater than $g(|w\rangle)$. Geometrically, QAS rotates the initial superposition $|\psi_0\rangle$ towards $\frac{1}{\sqrt{w+1}}\sum_{i=0}^{w}|i\rangle$, which is an average over the states that can reduce the state loss function from $|w\rangle$. If the output of the mth iteration is $|w_{new}\rangle = |i\rangle$ for some $i \geq w$, QAS will not update $|w\rangle$. On the other hand, if $i < w$, QAS will update $|w\rangle$ with $|w_{new}\rangle$ which is equivalent to descend $|w\rangle$ to a state with a smaller state loss function value. In Figure 2.3, we visually illustrate the mechanism of QAS when $p = 2$.

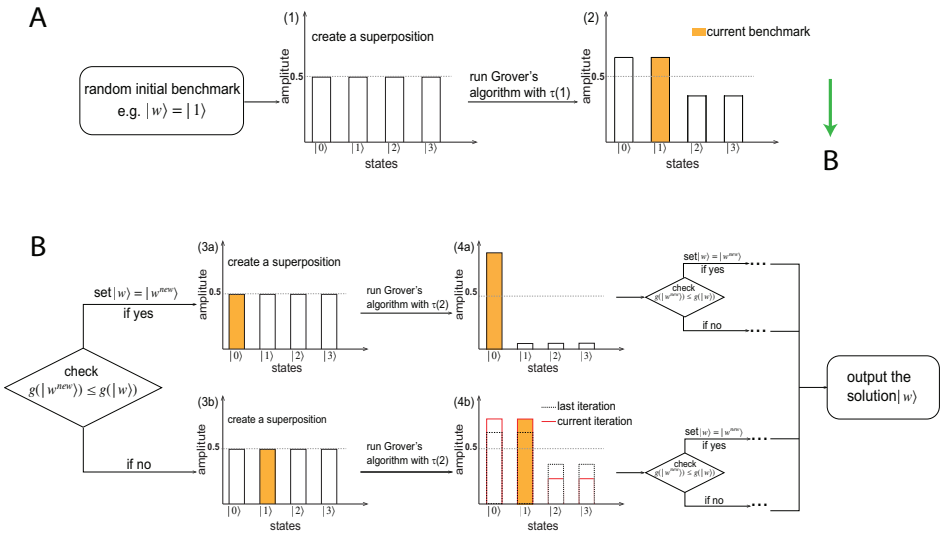

Figure 2.3: Illustrative example of quantum adaptive search.

Intuitively, one can think of QAS as a "quantum elevator" that starts at a random floor of a high tower and aims to descent to the ground floor. In each operation, a quantum machine will randomly decide if this elevator stays at the current floor or goes down to a random lower-level floor. Besides, the probability of going down will gradually increase as the number of operations increases. Thus, it is not hard to imagine that the "quantum elevator" can descent to the

ground floor with a high success probability after a large enough amount of operations. Our intuition of QAS is justified by the following theorem. Theorem 2.3.1 states that, within $O(\log_2 D)$ iterations, QAS finds the sole solution state for any pre-specified success probability greater than 50%.

Theorem 2.3.1. *Let $\kappa \in (0.5, 1)$ be a constant. With probability at least κ, the quantum adaptive search (e.g. Algorithm 2) finds the sole solution state $|0\rangle$ within $C_\kappa \log_2 D$ iterations, where C_κ is a positive constant that depends on κ.*

Suppose that, in the mth iteration of Algorithm 2, the current benchmark state is $|w_m\rangle$ with $g(|w_m\rangle$ being the r_mth smallest among $\{g(|i\rangle)\}_{i=0}^{D-1}$. Theorem 2.3.2 below shows the expected number of Grover's operations that Algorithm 2 needs to update $g(|w_m\rangle)$ is on the order $O(\sqrt{D/r_m})$. This together with Theorem 2.3.1 imply the computational complexity upper bound of QAS is of order $O(\sqrt{D}\log_2 D)$ which is only a $\log_2 D$ factor larger than the theoretical lower bound of any oracular quantum search algorithm (Bennett et al., 1997). Notice that, QAS achieves a near optimal computational efficiency as oracular quantum search algorithms without using any oracle information.

Theorem 2.3.2. *Let $|w_m\rangle$ be the current benchmark state in the mth iteration of Algorithm 2. Let r_m be the rank of $g(|w_m\rangle)$ in the sorted sequence of $\{g|i\rangle\}_{i=0}^{D-1}$ in ascending order, $r_m = 1, \ldots, D$. The expected time for Algorithm 2 to update $|w_m\rangle$ is of order $O(\sqrt{D/r_m})$.*

2.4 Implementation of quantum adaptive search

2.4.1 Linear prediction with quantum computing

Linear regression is a pivotal component of QAS as the local evaluation function compares two states through the state loss function defined in (2.3). Recently, a line of studies (Rebentrost et al., 2014; G. Wang, 2017; Wiebe et al., 2012; Z. Zhao et al., 2019) focuses on formulating linear regression as a matrix inversion problem, which can be tackled by the quantum algorithm to solve linear systems of equations (Harrow et al., 2009). Their goal is to find a series of quantum states whose amplitudes represent the ordinary least squares estimator of linear regression coefficients. Although their algorithms can encode such quantum states efficiently, "it may be exponentially expensive to learn via tomography" (Harrow et al., 2009). As a result, most existing quantum linear regression algorithms require the input design matrix to be sparse. Otherwise, the computation is slow and the estimation accuracy may suffer from noise accumulation. In this chapter, we investigate a novel quantum linear prediction

approach which is based on the singular-value decomposition. The proposed method avoids the matrix inversion problem. Instead, we estimate the inverse singular values by utilizing a recently developed quantum tomography technique (Lloyd et al., 2014).

Let $\{j_1, \ldots, j_d\}$ be a subset of $\{1, \ldots, p\}$ with $d \leq n$. Denote $\boldsymbol{X}_d \in \mathbb{R}^{n\times d}$ the sub-matrix of $\boldsymbol{X}$ corresponding to the subset $\{j_1, \ldots, j_d\}$. Let $\widetilde{\boldsymbol{x}}$ be a new observation of the d selected covariates. Then, the corresponding response value $\widetilde{y}$ can be predicted by

$$\widehat{y} = \widetilde{\boldsymbol{x}}^\intercal (\boldsymbol{X}_d^\intercal \boldsymbol{X}_d)^{-1} \boldsymbol{X}_d^\intercal \boldsymbol{y}. \tag{2.5}$$

Suppose that the rank of $\boldsymbol{X}_d$ is $R \leq d$, a compact singular value decomposition of $\boldsymbol{X}_d$ yields

$$\boldsymbol{X}_d = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal = \sum_{r=1}^{R} \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^\intercal,$$

where $\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_1, \ldots, \sigma_R\}$ is a diagonal matrix of R positive singular values, and $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_R) \in \mathbb{R}^{n\times R}$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_R) \in \mathbb{R}^{d\times R}$ are matrices of left and right singular vectors, respectively. Further, we have $\boldsymbol{U}\boldsymbol{U}^\intercal = \boldsymbol{V}\boldsymbol{V}^\intercal = \boldsymbol{I}_R$. Then, the linear predictor in (2.5) can be represented by

$$\widehat{y} = \sum_{r=1}^{R} \sigma_r^{-1} \widetilde{\boldsymbol{x}}^\intercal \boldsymbol{v}_r \boldsymbol{u}_r^\intercal \boldsymbol{y}. \tag{2.6}$$

Motivated by the quantum principle component analysis (QPCA) (Lloyd et al., 2014) and the inverted singular value problem (Schuld et al., 2016), we propose to calculate (2.6) by a quantum linear prediction (QLP) procedure, which can be summarized as the following three steps.

Step 1: Quantum states preparation.

We encode $\boldsymbol{X}_d$, $\boldsymbol{v}_r$ and $\boldsymbol{u}_r$, $r = 1, \ldots, R$, into a quantum system as

$$|\psi_{\boldsymbol{X}_d}\rangle = \sum_{j=1}^{d}\sum_{i=1}^{n} a_{i,j} |i\rangle |j\rangle, \quad |\boldsymbol{u}_r\rangle = \sum_{i=1}^{n} \mu_{r,i} |i\rangle \quad \text{and} \quad |\boldsymbol{v}_r\rangle = \sum_{j=1}^{d} \nu_{r,j} |j\rangle, \tag{2.7}$$

where $\sum_i \sum_j |a_{i,j}|^2 = \sum_i |\mu_{r,i}|^2 = \sum_j |\nu_{r,j}|^2 = 1$, and $\{|i\rangle\}_{i=1}^n$ and $\{|j\rangle\}_{j=1}^d$ are orthonormal quantum bases representing the linear space spanned by the left and right singular vectors of $\boldsymbol{X}_d$, respectively.

Then, we encode $\boldsymbol{X}_d^\intercal \boldsymbol{X}_d$ and the compact singular value decomposition of $\boldsymbol{X}_d$ into a quantum system as

$$\rho \equiv \rho(\boldsymbol{X}_d^\intercal \boldsymbol{X}_d) = \sum_{j,j'=1}^{d} \sum_{i=1}^{n} a_{i,j} a_{i,j'} \left|j\right\rangle \left\langle j'\right| \quad \text{and} \quad \left|\psi_{\boldsymbol{X}_d}\right\rangle = \sum_{r=1}^{R} \sigma_r \left|\boldsymbol{u}_r\right\rangle \left|\boldsymbol{v}_r\right\rangle . \tag{2.8}$$

Similarly, we can encode $\boldsymbol{y}$ and $\widetilde{\boldsymbol{x}}$ into a quantum system as

$$\left|\psi_{\mathbf{y}}\right\rangle = \sum_{\mu=1}^{N} b_\mu \left|\mu\right\rangle \quad \text{and} \quad \left|\psi_{\widetilde{\boldsymbol{x}}}\right\rangle = \sum_{\gamma=1}^{d} c_\gamma \left|\gamma\right\rangle , \tag{2.9}$$

with $\sum_\mu |b_\mu|^2 = \sum_\gamma |c_\gamma|^2 = 1$, and $\{|\mu\rangle\}_{\mu=1}^n$ and $\{|\gamma\rangle\}_{\gamma=1}^d$ are orthonormal quantum bases.

Step 2: Extracting the inverted singular values.

To calculate (2.6) with a quantum computer, we first need to calculate the inverse of singular values of $\boldsymbol{X}_d$. According to the ideas of QPCA (Lloyd et al., 2014), we apply ρ to $|\psi_{\boldsymbol{X}_d}\rangle$ and follow the quantum phase estimation algorithm (e.g. Shor, 1994; Szegedy, 2004; Wocjan et al., 2009) which yields

$$\sum_{r=1}^{R} \sigma_r \left|\boldsymbol{v}_r\right\rangle \left|\boldsymbol{u}_r\right\rangle \left|\lambda_r\right\rangle , \tag{2.10}$$

where $\lambda_r = \sigma_r^2$, $r = 1, \ldots, R$, are the eigenvalues of ρ which is encoded in a quantum register $|\lambda_r\rangle$. Then, by adding an extra qubit and rotating (2.10) conditional on the eigenvalue register $|\lambda_r\rangle$, we have

$$\sum_{r=1}^{R} \sigma_r \left|\boldsymbol{v}_r\right\rangle \left|\boldsymbol{u}_r\right\rangle \left|\lambda_r\right\rangle \left[\sqrt{1 - (\frac{c}{\lambda_r})^2}\left|0\right\rangle + \frac{c}{\lambda_r}\left|1\right\rangle\right] , \tag{2.11}$$

where c is a constant to ensure the inverse of eigenvalues are no larger than 1.

We can repeatedly perform a conditional measurement of (2.11) on the ancilla qubit until it is in state $|1\rangle$. After that, we can uncompute and discard the eigenvalue register which results in

$$\left|\psi_1\right\rangle \equiv \frac{1}{\sqrt{p(1)}} \sum_{r=1}^{R} \frac{c}{\sigma_r} \left|\boldsymbol{v}_r\right\rangle \left|\boldsymbol{u}_r\right\rangle , \tag{2.12}$$

where $p(1) = \sum_{r=1}^{R} |\frac{c}{\lambda_r}|^2$ is the probability of the ancilla qubit in (2.11) collapsing to state $|1\rangle$.

Step 3: Calculating the inner products.

Follow the notations in (2.7) –(2.9), we can rewrite (2.6) as

$$\sum_{r=1}^{R} \sigma_r^{-1} \langle \widetilde{\boldsymbol{x}} | \boldsymbol{v}_r \rangle \langle \boldsymbol{y} | \boldsymbol{u}_r \rangle . \tag{2.13}$$

Motivated by the strategy in (Schuld et al., 2016), we write (2.13) into some entries of an ancilla qubit so that it can be assessed by a single measurement. To this end, we construct two states $|\psi_1\rangle$ and $|\psi_2\rangle$ that are entangled with an ancilla qubit, i.e.

$$\frac{1}{\sqrt{2}} \left(|\psi_1\rangle |0\rangle + |\psi_2\rangle |1\rangle \right),$$

where $|\psi_1\rangle$ is defined in (2.12) and $|\psi_2\rangle = |\psi_{\boldsymbol{y}}\rangle |\psi_{\widetilde{\boldsymbol{x}}}\rangle$. Then, the off-diagonal elements of this ancilla qubit's density matrix read

$$\frac{c}{2\sqrt{p(1)}} \sum_{r=1}^{R} \sigma_r^{-1} \langle \widetilde{\boldsymbol{x}} | \boldsymbol{v}_r \rangle \langle \boldsymbol{y} | \boldsymbol{u}_r \rangle = \frac{c}{2\sqrt{p(1)}} \sum_{r=1}^{R} \sigma_r^{-1} \widetilde{\boldsymbol{x}}^\intercal \boldsymbol{v}_r \boldsymbol{u}_r^\intercal \boldsymbol{y},$$

which contains the desired results (2.6) up to a known normalization factor.

Given a testing set of n_t observations, i.e. $\{\widetilde{y}_i, \widetilde{\boldsymbol{x}}_i\}_{i=1}^{n_t}$, the prediction error can be calculated as

$$\tilde{g}(|i\rangle) = \frac{1}{n_t} \sum_{i=1}^{n_t} (\widehat{y}_i - \widetilde{y}_i)^2, \tag{2.14}$$

where $|i\rangle$ is a quantum state represents $\{x_1, \ldots, x_d\}$ and $\widehat{y}_i$ is the predictor of $\widetilde{y}_i$ which can be calculated follow the QLP procedure introduced above. In addition, the computational complexity of running QLP over a training set of size n and a testing set of n_t is approximately of the order $O\left((n + n_t) \log_2 d\right)$ which is faster than the classical computing algorithm that usually of order $O\left((n + n_t)d\right)$.

2.4.2 Hybrid quantum-classical strategy

In high dimensional regime, implementing the best subset selection procedure completely on a quantum computing system is desirable since both QAS and QLP are substantially faster than their classical counterparts. However, such an objective may not be readily accomplished due to the prototypical development of quantum computers. On one hand, the capacity of the state-of-the-art

quantum processor (about 70 qubits) is still far from large enough to carry out big data applications. For example, the quantum state preparation step of QLP requires at least $2(\log_2 d + \log_2 n)$ qubits which may exceed the capacity of most public available quantum computers when d and n are moderate or large. On the other hand, existing quantum computing systems are usually highly sensitive to the environment and can be influenced by both internal and external noises (Sung et al., 2019). For example, a superconducting quantum computing system can be affected by the internal noises due to material impurities as well as the external noises that come from control electronics or stray magnetic fields (Kandala et al., 2019). Further, such noises can accumulate through the computing process and create a non-negligible bias.

To address the aforementioned practical challenges, we propose a hybrid quantum-classical strategy to balance computational efficiency, scalability, and reliability of quantum best subset selection, which is summarized in Algorithm 3 below.

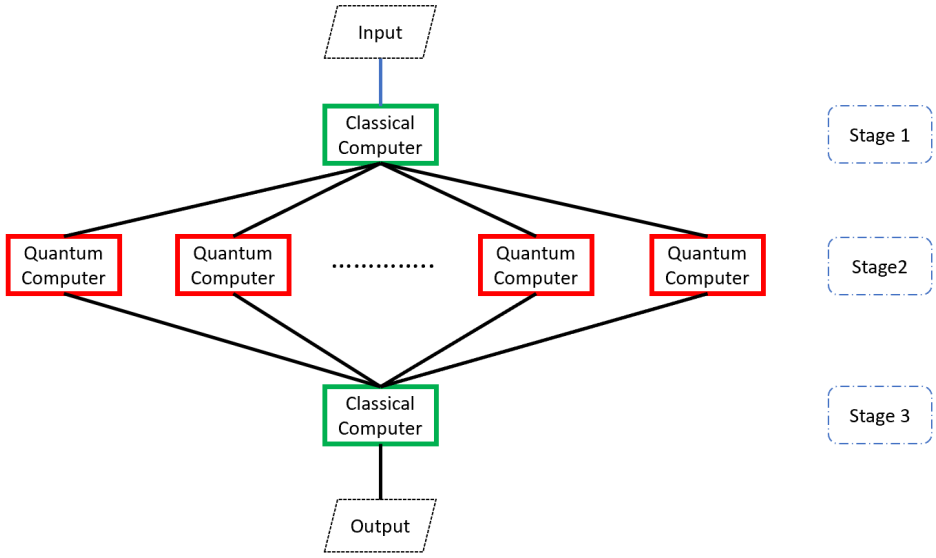

Figure 2.4: Flowchart of hybrid quantum-classical strategy.

The intuition of the hybrid quantum-classical strategy is to avoid the NP-hard computational bottleneck by implementing QAS on several quantum nodes while calculating the remaining parallelable steps on a classical computer. The accuracy of QAS will be boosted by a majority voting step. To be specific, we first input the data into a classical computer and parallelly compute the linear prediction errors (2.14) over $D = 2^p$ candidate models. The linear prediction errors, stored in a D-dimensional vector, will be passed to K quantum nodes.

Then, we independently implement QAS on K quantum nodes to select the minimum element in the D-dimensional prediction error vector. The selection results, stored in a list of K items, will be passed back to the classical computer for a majority voting. The best subset with the most votes is selected as the final estimator. We present the flow chart of this procedure in Figure 2.4 above.

Algorithm 3 Hybrid quantum-classical strategy for best subset selection

Input: A training set $\{\boldsymbol{x}_i,\ y_i\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^p$. A testing set $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^{n_t}$ with $\widetilde{\boldsymbol{x}}_i \in \mathbb{R}^p$. A number of available quantum nodes K. A learning rate $\lambda \in (0, 1)$.

Stage 1 on a classical computer:

(1.1) Compute, in parallel, the prediction error (2.14) over all $D = 2^p$ candidate models.

(1.2) Save the prediction errors $\boldsymbol{G}_D = (g(|0\rangle),\ \ldots,\ g(|D-1\rangle))$.

Stage 2 on K quantum nodes:

(2.1) Independently implement Algorithm 2 over $\boldsymbol{G}_D$ on K quantum nodes.

(2.2) Save the results in a list of K selected models $\mathcal{M} = \{m_1,\ \ldots,\ m_K\}$.

Stage 3 on a classical computer:

A majority voting over models in $\mathcal{M}$. Denote the model with the most votes as $\widehat{m}$.

Output: The selected model $\widehat{m}$.

The Theorem 2.4.1 below states a majority voting over $K = 2\xi + 1$ independent quantum nodes can boost the success probability of finding the best subset. The lower bound of success probability in Theorem 2.4.1 can be boosted by increasing the number of quantum nodes or improving the success probability on each quantum node.

Theorem 2.4.1. *Let's consider a majority voting system that consists of $K = 2\xi + 1$ nodes, where ξ is a positive integer. Suppose that each node works independently with probability $q > \frac{\xi+1}{2\xi+1}$ to vote the correct model and probability $1 - q$ to vote an incorrect model. Let $\mathcal{E}$ denote the event that the majority voting system selects the correct model. Then, the probability of $\mathcal{E}$ is lower bounded by*

$$\mathbb{P}(\mathcal{E}) > \Phi\left(\sqrt{2(2\xi+1)D_{KL}(q, \frac{\xi+1}{2\xi+1})}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable and

$$D_{KL}(a, b) = b \ln \frac{b}{a} + (1-b) \ln \frac{(1-b)}{(1-a)}$$

is the Kullback-Leibler divergence between two Bernoulli distributions with parameters a and b, respectively.

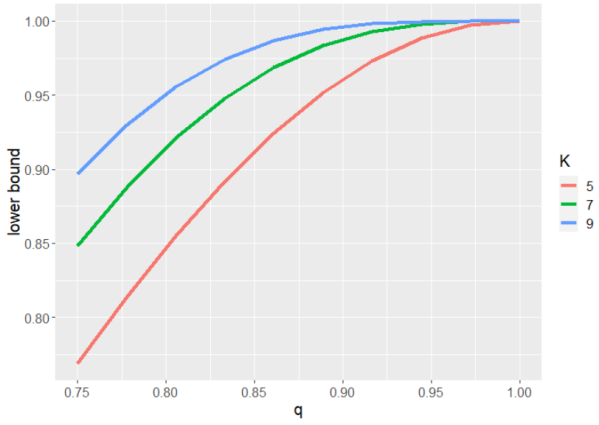

Figure 2.5: Lower bounds of successful probability when $K = 5, 7, 9$.

In Figure 2.5, we plot the the lower bound of $\mathbb{P}(\mathcal{E})$ versus the values of q. We let q increase from 0.75 to 1 and choose the number of nodes $K = 5, 7, 9$. According to Figure 2.5, all three solid lines are concave curves which means a majority voting system with a moderate number of nodes can effectively boost the success probability of finding the correct model. For instance, a majority voting over $K = 9$ independent nodes with $q = 0.75$ can boost the probability of finding the correct model to 0.9 or higher.

2.5 Experiments on the quantum computer

In this section, we implement the proposed hybrid quantum-classical strategy on IBM Quantum Experience[1] which is a publicly available cloud-based quantum computing system. This platform has developed a Qiskit Python development kit[2], which allows users to perform both quantum computing and classical computing in a single project. In Section 2.5.1, we analyze the performance of Algorithm 3 regarding the selections of tuning parameters. In Section 2.5.2, we assess the best subset section performance of Algorithm 3 under various linear regression settings and compare QAS with Grover's algorithm in stage 2.

[1] www.ibm.com/quantum-computing

[2] https://qiskit.org/

2.5.1 Selection of tuning parameters

The proposed hybrid quantum-classical strategy involves two tuning parameters: the number quantum nodes K; and a learning rate $\lambda \in (0, 1)$ which will

be passed to each quantum node to implement Algorithm 2. As suggested by Theorem 2.4.1, the success probability of Algorithm 3 converges to 1 as K increases when the success probability of each individual quantum node surpasses 0.5. Therefore, one should choose K as large as possible subject to the availability of quantum computing resources. On the other hand, the learning rate λ controls the "step size" of each iteration in Algorithm 2. In this subsection, we use an experiment to assess the performance and the sensitivity of Algorithm 3 with respect to the choices of K and λ. To focus on the selection of tuning parameters, we skip the stage 1 in Algorithm 3. Instead, we generate $\boldsymbol{G}_D$ as an i.i.d. sample of size $D = 32$ from the uniform distribution $U[0, 1]$. Then we use the stages 2 and 3 in Algorithm 3 to find the minimum element in $\boldsymbol{G}_D$. We set $K = 1, 3$ or 5, and let λ be a sequence of grid points between 0.40 and 0.60 with a step size of 0.01. For each pair of K and λ, we simulate 200 replications. The performance is measured by the accuracy rate which is defined as

$$\text{Accuracy rate} = \frac{\text{Number of replications with correct solution}}{\text{Number of replications}}. \tag{2.15}$$

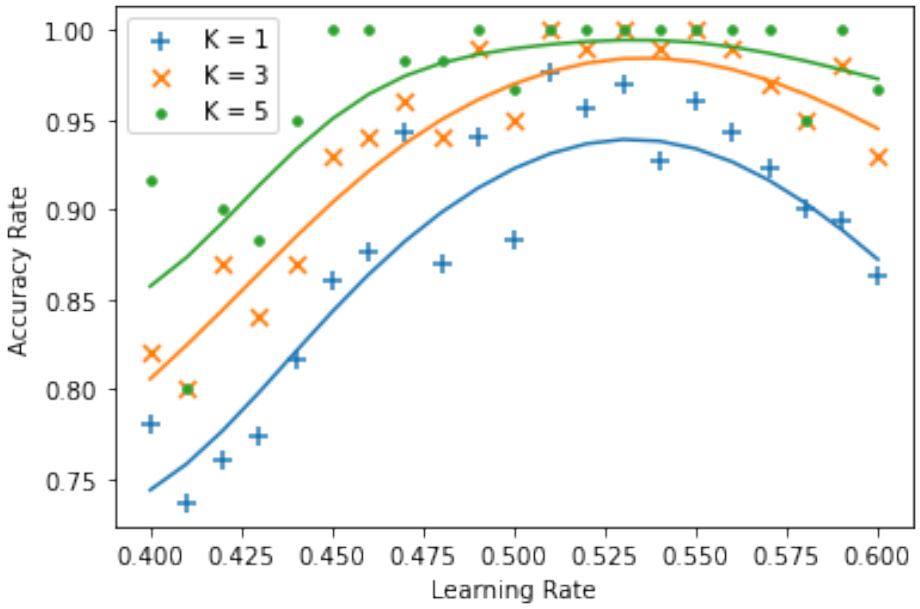

Figure 2.6: Accuracy rates of Algorithm 3 with $K = 1, 3, 5$ and $\lambda \in [0.40, 0.60]$.

The results are presented in Figure 2.6. The accuracy rates for $K = 1, 3$ or 5 are plotted as blue, orange and green dots with different symbols. The solid lines are corresponding smoothed curves. According to Figure 2.6, we find that increasing the number of quantum nodes K can improve the accuracy rate though the improvement is marginal when K is large enough, e.g. $K = 5$. Besides, we observe that the highest accuracy rates are attained with some λ between 0.5 and 0.55 for both choices of K. One can notice that the accuracy rates are close to 1 when $\lambda \in (0.5, 0.55)$ and $K \geq 3$ which can be considered as a rule-of-thumb recommendation. Further, the smoothed curves in Figure 2.6 indicates that Algorithm 3 is not very sensitive for the selection of tuning parameters.

2.5.2 Best subset selection with hybrid quantum-classical strategy

In this subsection, we assess the best subset selection performance of the proposed hybrid quantum-classical strategy. We consider a linear regression model

$$y_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}^* + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.16}$$

where $y \in \mathbb{R}$, $\boldsymbol{x}_i \in \mathbb{R}^p$, $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_p)^\intercal \in \mathbb{R}^p$, and $\epsilon_i \in \mathbb{R}$. First, we draw $\{\boldsymbol{x}_i\}_{i=1}^n$ as an i.i.d. sample from a multivariate normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where the (i, j)th entry of $\boldsymbol{\Sigma}$ equals $\rho^{|i-j|}$ for some $\rho \in (0, 1)$. Then, we draw $\{\epsilon_i\}_{i=1}^n$ as an i.i.d. sample from a normal distribution $N(0, \sigma^2)$. Notice that, ρ controls the autocorrelation level among covariates, and $\boldsymbol{\beta}^*$, ρ and σ together control the signal to noise ratio (SNR), i.e. SNR $= \boldsymbol{b}^{*\intercal} \boldsymbol{\Sigma} \boldsymbol{b}^* / \sigma^2$. Let $s < p$ be a positive integer indicating the model sparsity. We set the first s elements of $\boldsymbol{\beta}^*$ to be 1 and the rest $p - s$ elements to be 0.

Subject to the capacity of the quantum computing system, we set $n = 100$, $p = 7$, and $s = 4$. We choose the autocorrelation level $\rho \in \{0.25, 0.5\}$ and the signal to noise ratio SNR $\in \{0.5, 1.0, 2.0, 3.0\}$, respectively. For each scenario, we simulate 200 replications. We implement the hybrid quantum-classical strategy as proposed in Algorithm 3 with $K = 3$ and $\lambda = 0.55$ to select the best subset of the linear regression model. We denote this method as QAS since it implements the quantum adaptive search in stage 2. Besides, we consider two variants of Algorithm 3 as two competitors. The first competitor, denoted as Grover Oracle, replace QAS in stage 2 by Grover's algorithm with a known oracle state. The second competitor, denoted as Grover Random, replace QAS in stage 2 by Grover's algorithm with a randomly selected

oracle state. Notice that Grover Oracle is an oracular method and not applicable in practice since the oracle state is not observable.

For each replication, we record the false positives (FP) and false negatives (FN) of each method. FP is the number of inactive covariates that are selected into the model and FN is the number of active covariates that are ignored in the selected model. The box-plots of FP and FN over 200 replications are reported in Figure 2.7 below. According to Figure 2.7, Grover Oracle performs the best among the three competitors which is not surprising as it utilizes the oracle information of the true best subset. The proposed QAS method, as a non-oracular approach, performs almost identical to Grover Oracle in nearly all scenarios. QAS is slightly outperformed by Grover Oracle when the correlation level is high ($\rho = 0.5$) and the signal to noise ratio is small (SNR $= 0.5$), which is the most challenging scenario. In contrast, Grover Random performs as bad as random guesses in all scenarios. This observation, which is inline with the discussions in Section 2.2.3, shows oracular quantum search algorithms are not applicable to statistical learning problems.

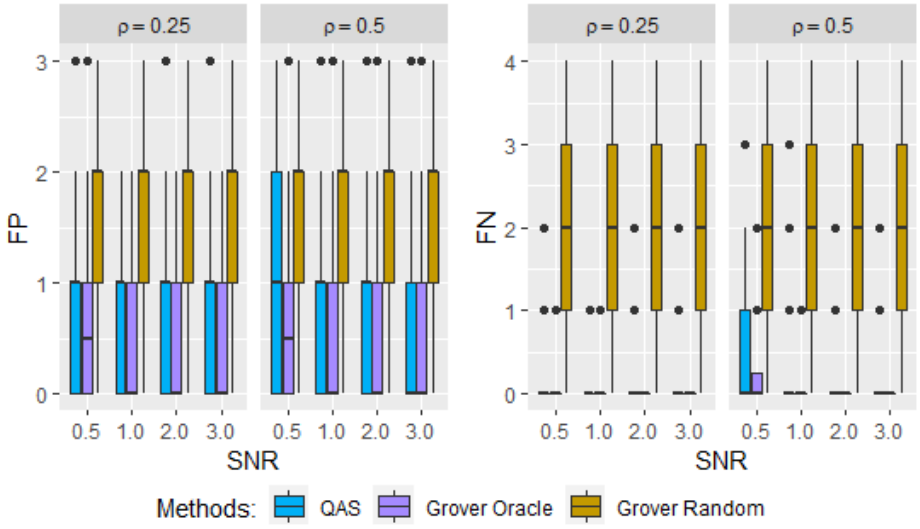

Figure 2.7: Box-plots of false positives (left) and false negatives (right) over 200 replications.

2.6 Comparison with classical methods

In this section, we compare the empirical performance of quantum adaptive search (QAS) with the naive best subset selection method (BSS) implemented

on a classical computer. We also include some popular approximate best subset selection methods designed for classical computers: FORWARD STEPWISE selection (e.g. Draper & Smith, 1966); LASSO (Tibshirani, 1996); and the fast best subset selection method (LOLEARN) (Hazimeh & Mazumder, 2018). The implementation details of the 5 comparison methods are summarized in Table 2.1.

Table 2.1: Implementation details for the 5 comparison methods in Section 2.6.

Method name	Computing method	Tuning parameter selection
QAS	Algorithm 3	$K = 5$ and $\lambda = 0.5$
BSS	Naive best subset selection	NA
FORWARD STEPWISE	R package bestsubset	NA
LASSO	R package glmnet	10-folds cross-validation
LOLEARN	R package LoLearn	10-folds cross-validation

We consider the same linear regression model in (2.16). The coefficient vector $\boldsymbol{\beta}^*$ is generated from one of the following two settings.

(1) **Strong sparsity:** the first s elements of β^* equal to 1 while the rest $p - s$ elements equal to 0.

(2) **Weak sparsity:** the first s elements of β^* equal to $1, \frac{s-1}{s}, \frac{s-2}{s}, \ldots, \frac{1}{s}$ while the rest $p - s$ elements equal to 0.

In this experiment, we set $n = 100$, $p = 10$ and $s = 5$. We choose the autocorrelation level $\rho \in \{0.25, 0.5\}$ and the signal to noise ratio SNR $\in \{0.5, 1.0, 2.0, 3.0\}$, respectively. For each scenario, we simulate 100 replications. Additional results for $p = 20$ and a real data analysis are relegated to the supplemental material. Besides the false positives (FP) and false negatives (FN) defined in Section 2.5.2, we measure the prediction performance of each method by the relative test error (RTE):

$$\text{RTE} = \mathbb{E}(\widetilde{y} - \widetilde{\boldsymbol{x}}^T\widehat{\boldsymbol{\beta}})^2/\sigma^2 = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\intercal\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)/\sigma^2 + 1,$$

where $(\tilde{y}, \tilde{\boldsymbol{x}})$ is a testing observation which is i.i.d. with the training sample $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$. It is easy to see that RTE is lower bounded by 1 and the smaller the better.

Figures 2.8 and 2.9 present box-plots of FP, FN and RTE over 100 replications under strong sparsity and weak sparsity settings, respectively. According to the box-plots, none of the 5 competing methods dominate the others in terms of all three measurements, which is inline with the discussions in (Hastie et al.,

2020). Generally speaking, LASSO tends to select generous models which may have small FNs but large FPs. FORWARD STEPWISE and LOLEARN prefer to select parsimonious models which may lead to small FPs but large FNs. However, none of the three methods can be considered as a good approximation of BSS since their performance are distinct in most scenarios. In contrast, QAS performs almost identical to BSS in terms of all three measurements in every scenario.

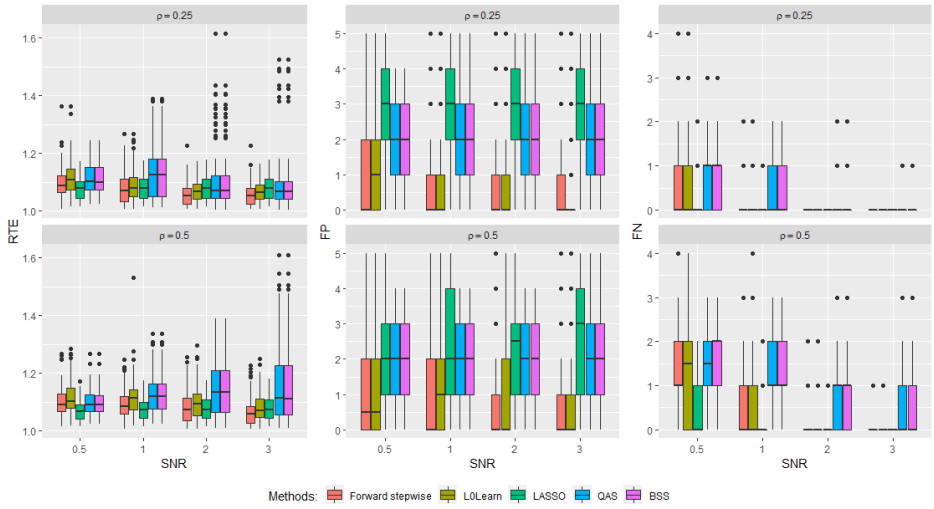

Figure 2.8: **Strong sparsity setting:** Boxplots of relative test error (left), false positives (middle) and false negatives (right) over 100 replications.

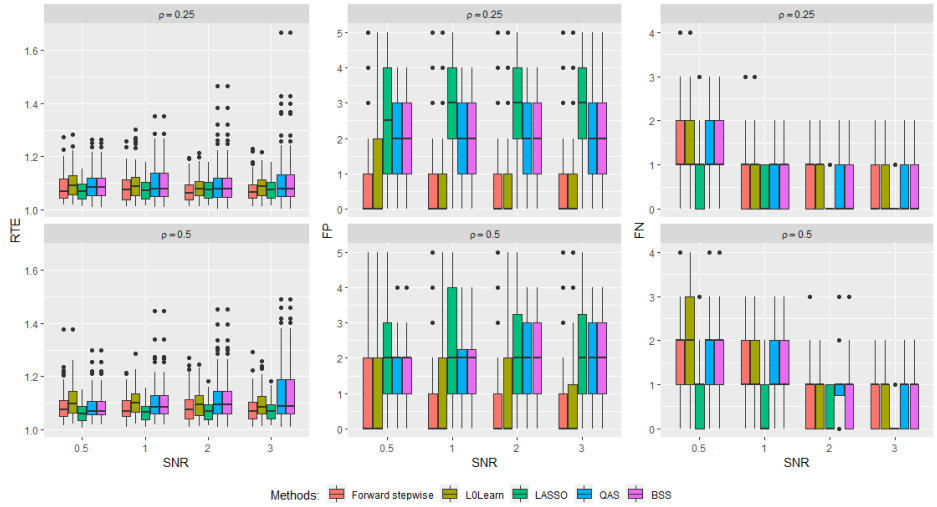

Figure 2.9: **Weak sparsity setting:** Boxplots of relative test error (left), false positives (middle) and false negatives (right) over 100 replications.

Next, we take a closer look at best subset selection behaviors of the 5 comparison methods. We simulate 1,000 replications of (2.16) with $n = 100$, $p = 10$, $s = 5$, $\rho = 0.5$, SNR $= 0.5$, and the weak sparsity $\boldsymbol{\beta}^*$. In the top panel of Figure 2.10, we present the histogram of selected model sizes for 5 methods. In the bottom panel of Figure 2.10, we report the box-plots of RTSs for 5 methods. According to Figure 2.10, the selected model sizes of QAS and BSS sharply concentrates around the truth, i.e. $s = 5$. The histograms of FORWARD STEPWISE and LOLEARN are severely left skewed which indicates they tend to underestimate the size of the active set. In contrast, LASSO has a right skewed histogram and hence often select oversized models. Again, the selection as well as prediction performances of QAS and BSS are almost identical. The results in Figure 2.10 further justified our argument that QAS is an efficient quantum alternative of BSS while FORWARD STEPWISE, LASSO and LOLEARN are not ideal approximates of BSS.

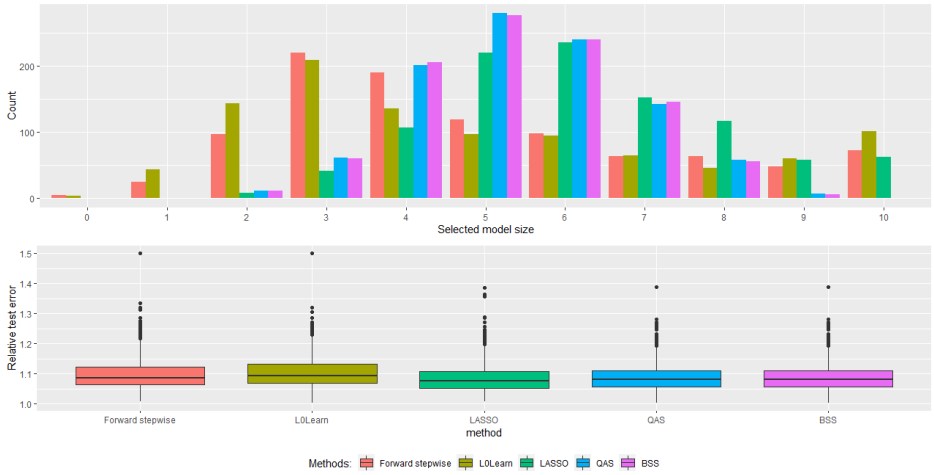

Figure 2.10: **Best subset selection behaviors:** histogram of selected model size (top) and boxplots of relative test error (bottom).

2.7 Additional numerical results

2.7.1 Additional results for Section 6

In this subsection, we report additional simulation results for Section 6. All the simulation settings are the same as introduced in Section 6 except that we increase p to 20. Figures 2.11 and 2.12 present box-plots of FP, FN and RTE over 100 replications under strong sparsity and weak sparsity settings, respectively. The results for $p = 20$ are similar as the results we reported in the main document for $p = 10$.

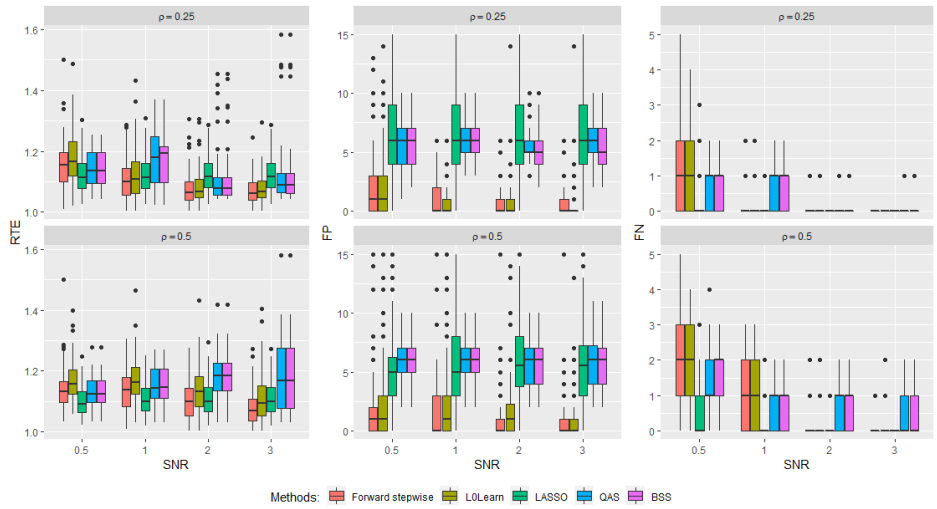

Figure 2.11: **Strong sparsity setting with** $p = 20$**:** Boxplots of relative test error (left), false positives (middle) and false negatives (right) over 100 replications.

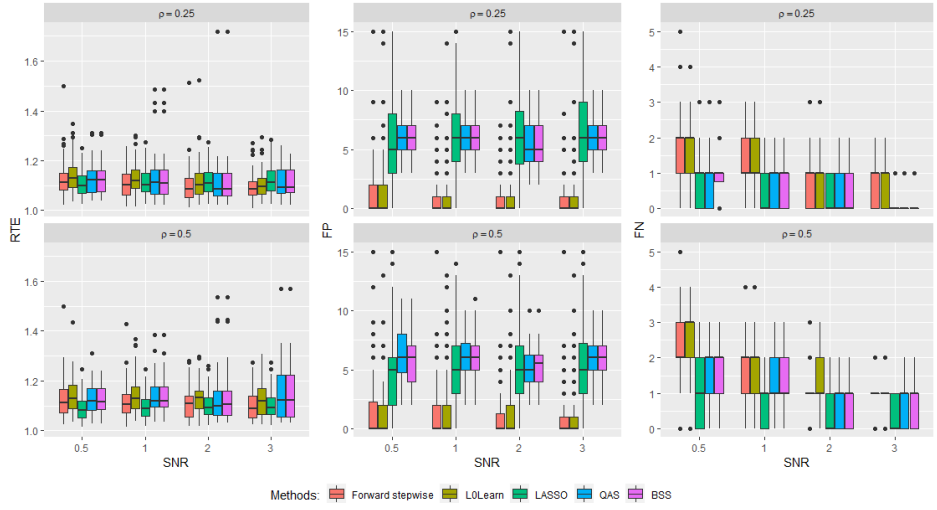

Figure 2.12: **Weak sparsity setting with** $p = 20$**:** Boxplots of relative test error (left), false positives (middle) and false negatives (right) over 100 replications.

2.8 Real data analysis

We compare the performance of the methods listed in Table 1 on a real dataset. The dataset (Johnson, 1996) contains 18 measurements including body fat, age,

weight, height, and ten body circumference measurements, e.g., abdomen, on 252 men. The percentage of body fat (*brozek*) is considered as the response. More details are summarized in Table 2.2.

Table 2.2: Descriptions of 18 health measurements in the dataset (Johnson, 1996).

variable	**description**
brozek	Percent body fat using Brozek's equation
density	Density (gm/cm^3)
weight	Weight(lbs)
adipos	Adiposity index = Weight/Height^2 (kg/m^2)
neck	Neck circumference (cm)
abdom	Abdomen circumference (cm) at the umbilicus and level with the iliac crest
thigh	Thigh circumference (cm)
ankle	Ankle circumference (cm)
forearm	Forearm circumference (cm)
siri	Percent body fat using Siri's equation
age	Age(yrs)
height	Height(inches)
free	Fat Free Weight = (1 - fraction of body fat) * Weight, using Brozek's formula (lbs)
chest	Chest circumference (cm)
hip	Hip circumference (cm)
knee	Knee circumference (cm)
biceps	Extended biceps circumference (cm)
wrist	Wrist circumference (cm) distal to the styloid processes

The goal is to find a linear model that can accurately predict body fat (*brozek*) using the other variables except *siri* (another way of computing body fat), *density* (it is used in the brozek and siri formulas) and *free* (it is computed using brozek formula). So the total number of predictors is 14. We randomly split the dataset into 80% training and 20% testing for 100 replications. The out-of-sample mean-squared error (Test MSE) is used as a evaluation metric in this study. In addition, we report the support size, i.e., the number of non-zero coefficients.

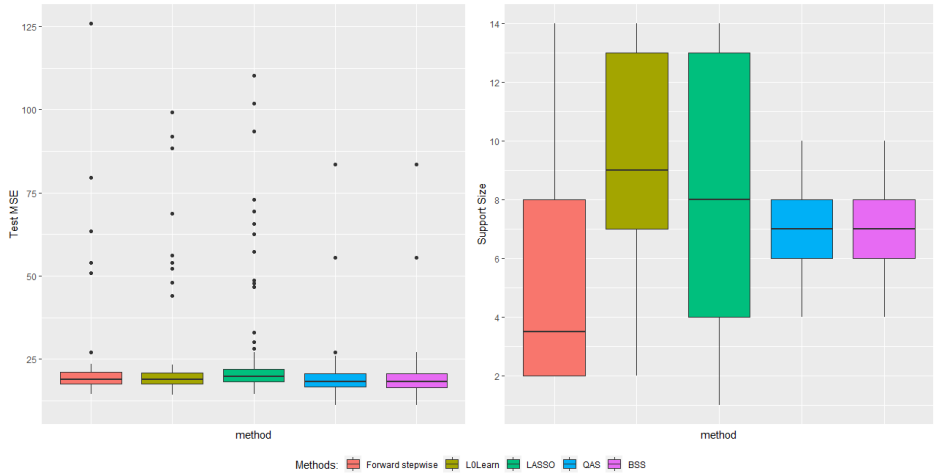

Figure 2.13: Left: boxplots of Test MSE for five methods. Right: boxplots of support sizes for five methods.

Table 2.3: Test MSEs and support sizes for five methods on the body fat dataset.

		Forward stepwise	LoLearn	LASSO	QAS	BSS
Test MSE	median	18.92	18.85	19.81	**18.12**	**18.12**
	mean	21.85	23.00	25.87	19.52	**19.51**
	sd	13.85	15.24	18.04	**8.06**	**8.06**
Support Size	median	3.50	9.00	8.00	**7.00**	**7.00**
	mean	**5.17**	9.49	8.21	6.99	6.99
	sd	3.43	3.44	4.36	**1.31**	**1.31**

According to Figure 2.13 and Table 2.3, we observe that QAS and BSS perform identical and outperform the other methods. Moreover, Table 2.4 shows the percentage of each predictor that has been selected within 100 repetitions across five different methods. According to Table 2.4, we can see *abdom* is the key factor in predicting the percentage of body fat. This finding is inline with the studies in (Ochiai et al., 2010) and (Flegal et al., 2009).

Table 2.4: Measurements selected percentage

Rank	Forward Stepwise	LoLearn	LASSO	QAS	BSS
1	abdom(100%)	abdom(100%)	abdom(100%)	abdom(100%)	abdom(100%)
2	weight(74%)	wrist(92%)	wrist(98%)	wrist(92%)	wrist(92%)
3	wrist(67%)	age(78%)	age(97%)	weight(84%)	weight(84%)
4	neck(38%)	height(73%)	height(94%)	forearm(84%)	forearm(84%)
5	forearm(36%)	forearm(71%)	forearm(77%)	neck(74%)	neck(74%)
6	age(33%)	neck(60%)	neck(76%)	thigh(67%)	thigh(68%)
7	biceps(28%)	biceps(52%)	ankle(66%)	age(62%)	age(61%)
8	height(22%)	thigh(49%)	biceps(64%)	hip(54%)	hip(54%)
9	hip(22%)	knee(45%)	thigh(52%)	biceps(26%)	biceps(26%)
10	thigh(22%)	ankle(45%)	hip(49%)	height(18%)	height(18%)
11	adipos(21%)	hip(44%)	weight(48%)	chest(12%)	chest(12%)
12	chest(20%)	weight(41%)	knee(48%)	adipos(11%)	adipos(11%)
13	ankle(18%)	chest(40%)	adipos(40%)	knee(11%)	knee(11%)
14	knee(16%)	adipos(32%)	chest(40%)	ankle(4%)	ankle(4%)

2.9 Proofs

2.9.1 Technical Lemmas

In this subsection, we provide two technical lemmas to pave the way for the proof of main theorems. We omit the proof of Lemma 2.9.1 as it can be found in the literature.

Lemma 2.9.1 (c.f. Lemma 1 in Boyer et al., 1998)**.** *For any real numbers α and β and any positive integer m, we have*

$$\sum_{j=0}^{m-1} \cos(\alpha + 2\beta j) = \frac{\sin(m\beta)\cos(\alpha + (m-1)\beta)}{\sin\beta}.$$

In particular, when $\alpha = \beta$, the above equality can be simplified as

$$\sum_{j=0}^{m-1} \cos(\alpha + 2\beta j) = \frac{\sin(2m\alpha)}{2\sin\alpha}.$$

Lemma 2.9.2. *Let s_0 be the (unknown) number of solutions states among D states. Let θ be such that $\sin^2\theta = s_0/D$. Let γ be an arbitrary positive integer.*

Let j be an integer sampled uniformly between 0 and $\tau - 1$. If we observe the register after applying j Grover's operations to the initial state $|\psi_0\rangle = \sum_{i=0}^{D-1} \frac{1}{\sqrt{D}} |i\rangle$, the probability of obtaining a solution is exactly

$$P_\tau = \frac{1}{2} - \frac{\sin(4\tau\theta)}{4\tau \sin(2\theta)}. \tag{2.17}$$

Further, we have $P_\tau \geq \frac{1}{4}$ when $\tau \geq \frac{1}{\sin(2\theta)}$.

Proof of Lemma 2.9.2. The probability of finding one solution state among s_0 if we perform j Grover's operations is $s_0 k_j^2 = \sin^2((2j+1)\theta)$. It follows that the average success probability when $0 \leq j < \tau$ is chosen randomly is

$$\begin{aligned} P_\tau &= \sum_{j=0}^{\tau-1} \frac{1}{\tau} \sin^2((2j+1)\theta) \\ &= \frac{1}{2\tau} \sum_{j=0}^{\tau-1} \{1 - \cos((2j+1)2\theta)\} \\ &= \frac{1}{2} - \frac{\sin(4\tau\theta)}{4\tau \sin 2\theta}. \end{aligned}$$

If $\tau \geq \frac{1}{\sin(2\theta)}$, we complete the proof as the following inequality holds

$$\frac{\sin(4\tau\theta)}{4\tau \sin 2\theta} \leq \frac{1}{4\tau \sin 2\theta} \leq \frac{1}{4}.$$

□

2.9.2 Proof of Theorem 3.1

For the ease of presentation and without loss of generality, we re-numbering the states in descending order in this proof, i.e.

$$g(|0\rangle) > g(|1\rangle) \geq g(|2\rangle) \geq \cdots \geq g(|D-1\rangle) > g(|D\rangle), \tag{2.18}$$

where $|D\rangle$ is the unique solution state and $|0\rangle$ is an added initial state to facilitate the discussion with $g(|0\rangle)$ being an arbitrarily large value.

Let's assume Algorithm 2 is initialized with the least favorable state $|0\rangle$. Let Z denote the number of iterations the algorithm takes to arrive at the solution state $|D\rangle$. The rule that Algorithm 2 moves from $|0\rangle$ to $|D\rangle$ can be abstracted as the following mathematical process.

> Iteration 1: Draw an integer X_1 uniformly from 0 to D, then the algorithm moves from $|0\rangle$ to $|X_1\rangle$.

ITERATION 2: Draw an integer X_2 uniformly from 0 to $D - X_1$, then the algorithm moves from $|X_1\rangle$ to $|X_1 + X_2\rangle$.

$$\vdots$$

ITERATION z: Draw an integer X_z uniformly from 0 to $D - \sum_{i=1}^{z-1} X_i$, then the algorithm moves from $\left|\sum_{i=1}^{z-1} X_i\right\rangle$ to $|\sum_{i=1}^{z} X_i\rangle$. If $\sum_{i=1}^{z} X_i = D$, then the algorithms stops at $Z = z$ as it finds the solution state $|D\rangle$. Otherwise, the algorithm goes to the $(z+1)$th iteration.

As we can see, the total number of iterations Z is a discrete random variable that can take any positive integer values. To prove Theorem 1, it is equivalent to show that the κth quantile of the discrete random variable Z is upper bounded by $C_\kappa \ln D$ for a positive constant C_κ. The proof will be unveiled by three steps.

Step 1: A partial sum process.

To investigate the probability distribution of Z, we first study a partial sum of X_i which is defined as follows

$$S_z = \sum_{i=1}^{z} X_i \text{ for } z = 1, 2, \ldots.$$

Notice that $(S_z | S_{z-1} = s_{z-1}, \ldots, S_1 = s_1) \overset{\mathcal{D}}{=} (S_z | S_{z-1} = s_{z-1}) \sim \text{unif}\{s_{z-1}, D\}$. Also, we can show $\{S_z\}_{z=1,2,\ldots}$ is a submartingale, i.e. $\mathbb{E}[S_z | S_1, \ldots, S_{z-1}] = \mathbb{E}[S_z | S_{z-1}] \geq S_{z-1}$. Thus, $\{S_z\}_{z=1,2,\ldots}$ is a discrete-time Markov chain with a finite state space $\{0, 1, 2, \ldots, D\}$.

Moreover, the expectation of S_z satisfies

$$\begin{aligned}
\mathbb{E}[S_z] &= \mathbb{E}\left[\mathbb{E}[S_z | S_{z-1}]\right] = \mathbb{E}\left[\frac{S_{z-1} + D + 1}{2}\right] \\
&= \mathbb{E}[S_{z-1}]/2 + (D+1)/2 \\
&= \mathbb{E}[X_1]/2^{z-1} + (D+1)\big(1 - 1/2^{z-1}\big) \\
&= (D+1)/2^z + +(D+1)\big(1 - 1/2^{z-1}\big) \\
&= (D+1)\big(1 - 1/2^z\big).
\end{aligned}$$

Step 2: The probability mass function of Z.

Next, we derive the probability mass function of Z. The derivation is based on the Markov property of the partial sum process $\{S_z\}_{z=1,2,\ldots}$. Define a tran-

sition matrix P as

$$P = \begin{pmatrix} \frac{1}{D+1} & \frac{1}{D+1} & \frac{1}{D+1} & \cdots & \frac{1}{D+1} & \frac{1}{D+1} \\ 0 & \frac{1}{D} & \frac{1}{D} & \cdots & \frac{1}{D} & \frac{1}{D} \\ 0 & 0 & \frac{1}{D-1} & \cdots & \frac{1}{D-1} & \frac{1}{D-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Let π_z be a row vector with dimension $D+1$ that denotes the distribution of the chain $\{S_z\}_{z=1,2,...}$ at time z. By the definition of $\{S_z\}_{z=1,2,\ ...}$ and P, we have

$$\pi_{z+1} = \pi_z P, \quad \text{for} \quad z = 1, 2, \ \ldots.$$

Hence, we can write out π_1, π_2 and π_3 as

$$\pi_1 = \Big(\frac{1}{D+1}, \frac{1}{D+1}, \frac{1}{D+1}, \ \cdots, \frac{1}{D+1}, \frac{1}{D+1}\Big),$$

$$\pi_2 = \frac{1}{D+1}\Big(\frac{1}{D+1}, \sum_{i_1=D}^{D+1}\frac{1}{i_1}, \sum_{i_1=D-1}^{D+1}\frac{1}{i_1}, \ \cdots, \sum_{i_1=2}^{D+1}\frac{1}{i_1}, \sum_{i_1=1}^{D+1}\frac{1}{i_1}\Big), \quad \text{and}$$

$$\pi_3 = \frac{1}{D+1}\Big(\frac{1}{(D+1)^2}, \sum_{i_2=D}^{D+1}\frac{1}{i_2}\sum_{i_1=i_2}^{D+1}\frac{1}{i_1}, \sum_{i_2=D-1}^{D+1}\frac{1}{i_2}\sum_{i_1=i_2}^{D+1}\frac{1}{i_1}, .., \sum_{i_2=2}^{D+1}\frac{1}{i_2}\sum_{i_1=i_2}^{D+1}\frac{1}{i_1}, \sum_{i_2=1}^{D+1}\frac{1}{i_2}\sum_{i_1=i_2}^{D+1}\frac{1}{i_1}\Big)$$

In general, we can summarize the expression of π_z for $z = 1, 2, \ \ldots$ as

$$\pi_z = \Big(\mathbb{P}(S_z = 0), \mathbb{P}(S_z = 1), \mathbb{P}(S_z = 2), \ \ldots, \mathbb{P}(S_z = D-1), \mathbb{P}(S_z = D)\Big),$$

where

$$\mathbb{P}(S_z = j) = \frac{1}{D+1}\sum_{i_{z-1}=D+1-j}^{D+1}\frac{1}{i_{z-1}}\sum_{i_{z-2}=i_{z-1}}^{D+1}\frac{1}{i_{z-2}}\cdots\sum_{i_1=i_2}^{D+1}\frac{1}{i_1}, \quad \text{for} \quad j = 0, 1, \ \ldots, D.$$

Therefore, we can write out the probability mass function of $Z = z$ for $z = 1, 2, \ldots$ as

$$\mathbb{P}(Z = z) = \mathbb{P}\big(S_z = D\big) = \frac{1}{D+1} \sum_{i_{z-1}=1}^{D+1} \frac{1}{i_{z-1}} \sum_{i_{z-2}=i_{z-1}}^{D+1} \frac{1}{i_{z-2}} \cdots \sum_{i_2=i_3}^{D+1} \frac{1}{i_2} \sum_{i_1=i_2}^{D+1} \frac{1}{i_1}. \tag{2.19}$$

Notice that the last summation in (2.19), i.e. $\sum_{i_1=i_2}^{D+1} \frac{1}{i_1}$, is a finite partial sum of a harmonic series. Therefore, we can write it as

$$\sum_{i_1=i_2}^{D+1} \frac{1}{i_1} = \ln(D+1) + \eta_{D+1} - \ln(i_2) - \eta_{i_2} \asymp \ln(D+1) - \ln(i_2),$$

where $\eta_z \asymp \frac{1}{2z}$.

Let us move on to the next level of summation in (2.19), which is upper bounded by

$$\sum_{i_2=i_3}^{D+1} \frac{1}{i_2} \sum_{i_1=i_2}^{D+1} \frac{1}{i_1} \asymp \sum_{i_2=i_3}^{D+1} \frac{1}{i_2} \big[\ln(D+1) - \ln(i_2)\big] \lesssim \ln(D+1)\big[\ln(D+1) - \ln(i_3)\big]$$
$$\lesssim \ln^2(D+1),$$

and lower bounded by

$$\sum_{i_2=i_3}^{D+1} \frac{1}{i_2} \sum_{i_1=i_2}^{D+1} \frac{1}{i_1} \asymp \sum_{i_2=i_3}^{D+1} \frac{1}{i_2} \big[\ln(D+1) - \ln(i_2)\big] \asymp \ln(D+1)\big[\ln(D+1) - \ln(i_3)\big]$$
$$- \sum_{i_2=i_3}^{D+1} \frac{\ln(i_2)}{i_2} \gtrsim \ln^2(D+1) - \ln(D+1)\ln(i_3),$$

since $\sum_{i_2=i_3}^{D+1} \frac{\ln(i_2)}{i_2} \asymp \frac{1}{2}\ln^2(D+1) - \frac{1}{2}\ln^2(i_3)$.

Similarly, we can go through all the summations in (2.19) and show

$$\mathbb{P}(Z = 0) = \frac{1}{D+1} \quad \text{and} \quad \mathbb{P}(Z = z) \asymp \frac{\ln^{z-1}(D+1)}{D+1}. \tag{2.20}$$

Step 3: The κth quantile of Z.

With the results in (2.20), we summarize the cumulative distribution function of Z as

$$F_Z(z) = \mathbb{P}(Z \leq z) = \sum_{j=0}^{z} \mathbb{P}(Z = j) \asymp \frac{1}{D+1} + \sum_{j=1}^{z} \frac{\ln^{j-1}(D+1)}{D+1} \asymp \frac{\ln^{z-1}(D+1)}{D+1}.$$

Let $z = c\ln(D+1) + 1$ with a positive constant $c \in (0, 1)$, we have

$$F_Z(z) \asymp \frac{\ln^{z-1}(D+1)}{D+1} \asymp \frac{(D+1)^{c\ln\ln(D+1)}}{D+1} \asymp 1.$$

Therefore, for any $\kappa \in (\frac{1}{2}, 1)$, there exists a positive constant C_κ such that the κth quantile of Z is upper bounded by $C_\kappa \log_2 D$ which completes the proof.

2.9.3 Proof of Theorem 3.2

Suppose we are in the mth iteration of Algorithm 2. Let $|w_m\rangle$ be the current benchmark state and r_m be the rank of $g(|w_m\rangle)$ in the sorted sequence of $\{g|i\rangle\}_{i=0}^{D-1}$ in ascending order, $r_m = 1, \ldots, D$. Notice that, Algorithm 2 implements $\tau(m) = \lceil \pi\lambda^{-m/2}/4 \rceil \equiv \lceil \frac{\pi}{4}\gamma^m \rceil$ Grover's operations, where $\gamma = \lambda^{-1/2}$. We are interested in finding the expected number of Grover's operations to update $|w_m\rangle$.

Let θ be the angle such that $\sin^2\theta = r_m/D$. Let

$$\tau_m^* = \frac{1}{\sin(2\theta)} = \frac{D}{2\sqrt{(D-r_m)r_m}} < \sqrt{\frac{D}{r_m}}.$$

We say that Algorithm 2 reaches a phase transition for $|w_m\rangle$ if $\tau(s)$ exceeds $\tau^*(m)$ for some $s \geq m$.

The expected total number of Grover's operations needed to reach the phase transition for $|w_m\rangle$ is upper bounded by

$$\frac{\pi}{4} \sum_{j=0}^{\lceil \log_\gamma(\tau_m^*) \rceil} \gamma^j < \frac{\pi}{4} \frac{\gamma^m - 1}{\gamma - 1} < \frac{\tau_m^* + 1}{\gamma - 1}.$$

Thus, if the algorithm updates $|w_m\rangle$ before it reaches reach the phase transition, the expected number of Grover's operations is at most of the order $O(\tau_m^*)$, which is further upper bounded by $O(\sqrt{D/r_m})$.

If the phase transition for $|w_m\rangle$ is reached, as proved by Lemma 2.9.2, every new iteration of Algorithm 2 will be able to update $|w_m\rangle$ with a probability at least 1/4 since $\tau(s) \geq \tau_m^*$. Then, the expected iterations to update $|w_m\rangle$

after the phase transition is upper bounded by a positive constant. Hence, the expected number of additional Grover's operations needed to update $|w_m\rangle$ after the phase transition is upper bounded by $C\tau_m^*$ for some positive constant C. These two scenarios together completes the proof.

2.9.4 Proof of Theorem 2.4.1

Recall that the $K = 2\xi + 1$ nodes work independently and share the same "success" probability q. Let $B \equiv B(\xi, q)$ denote the random variable that indicates the number of nodes who vote for the correct model in the system. It is easy to see that B follows a binomial distribution with parameters $2\xi + 1$ and q.

Then, the probability of $\mathcal{E}$ follows

$$\mathbb{P}(\mathcal{E}) = \sum_{i=\xi+1}^{2\xi+1} p_B(i) = 1 - \sum_{i=0}^{\xi} p_B(i) = 1 - F_B(\xi),$$

where $p_B(i)$ and $F_B(a)$ are the probability mass function and the cumulative distribution function (CDF) of B, respectively.

Theorem 1 in Short, 2013 provides an upper bound for $F_B(\xi)$, i.e.,

$$F_B(\xi) < \Phi\left(\text{sign}(\xi + 1 - Kq)\sqrt{2KD_{KL}(q, \frac{\xi+1}{K})}\right),$$

where $\Phi(\cdot)$ is the CDF of a standard normal random variable and $\text{sign}(\cdot)$ is the sign function. Further

$$D_{KL}(a, b) = b \ln \frac{b}{a} + (1 - b) \ln \frac{(1 - b)}{(1 - a)}$$

is the Kullback-Leibler divergence between two Bernoulli distributions with parameters a and b, respectively.

Thus, we bound the probability of $\mathcal{E}$ from below,

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &> 1-\Phi\left(\text{sign}(\xi+1-Kq)\sqrt{2KD_{KL}(q,\frac{\xi+1}{K})}\right)\\ &=\Phi\left(\text{sign}(Kq-\xi-1)\sqrt{2KD_{KL}(q,\frac{\xi+1}{K})}\right)\\ &=\Phi\left(\sqrt{2(2\xi+1)D_{KL}(q,\frac{\xi+1}{2\xi+1})}\right),\end{aligned}$$

where the last equality uses the fact $K=2\xi+1$ and the condition $q>\frac{\xi+1}{2\xi+1}$.

2.9.5 Proof of query complexity of Grover algorithm

This proof is a geometric proof based on Figure 2.2. Consider a plane spanned by $|k\rangle$ and $|\zeta\rangle$. Let $|\zeta\rangle=\frac{1}{\sqrt{D-1}}\sum_{i\neq k}|i\rangle$ be the average of all non-solution states which is orthogonal to the solution state $|k\rangle$, where $D=2^p$.

In the first iteration, the reflection angle θ could be calculated using inner product

$$\begin{aligned}<\psi,\zeta> &= (\frac{1}{\sqrt{2^p}}\sum_{i=0}^{D-1}|i\rangle)(\frac{1}{\sqrt{2^p-1}}\sum_{i\neq k}|i\rangle)\\ &=\frac{1}{\sqrt{2^p}}\frac{1}{\sqrt{2^p-1}}(2^p-1)\\ &=\sqrt{\frac{2^p-1}{2^p}}\\ &=|\psi|\times|\zeta|\times cos\theta\\ &=1\times1\times cos\theta\\ &=cos\theta\end{aligned}$$

Therefore,

$$\begin{aligned}cos\theta&=\sqrt{\frac{2^p-1}{2^p}}\\ sin\theta&=\sqrt{\frac{1}{2^p}}\end{aligned}$$

When p is large, $\frac{1}{\sqrt{2^p}}$ is a extremely small value, so we have

$$\begin{aligned} sin\theta &= \theta \\ &= \sqrt{\frac{1}{2^p}} \end{aligned}$$

So

$$\theta = \sqrt{\frac{1}{2^p}}$$

After reflecting ψ about ζ, we further reflect the state vector about the original ψ. Starting from a state vector which has an angle of $\frac{\pi}{2} - \theta$, after each iteration, the state vector moves angle 2θ closer to $|k\rangle$. Suppose the number of iterations we conducted is τ, we have

$$(\tau + \frac{1}{2}) \times 2\theta = \frac{\pi}{2}$$

Therefore,

$$\begin{aligned} \tau &= \frac{\pi}{4\theta} - \frac{1}{2} \\ &\approx \frac{\pi\sqrt{2^p}}{4} - -\frac{1}{2} \\ &\approx \frac{\pi\sqrt{2^p}}{4} \end{aligned}$$

Therefore, the asymptotically the query complexity is $O(\sqrt{2^p}) = O(\sqrt{D})$, where $D = 2^p$

2.10 Chapter conclusion

In this chapter, we investigated the solution of the best subset selection problem in a quantum computing system. We proposed an adaptive quantum search algorithm that does not require the oracle information of the solution state and can significantly reduce the computational complexity of classical best subset selection methods. An efficient quantum linear prediction method was also discussed. Further, we introduced a novel hybrid quantum-classical strategy to avoid the hardware bottleneck of existing quantum computing systems and boost the success probability of quantum adaptive search. The quantum adaptive search is applicable to various best subset selection problems. Moreover, our

study opens the door to a new research area: reinvigorating statistical learning with powerful quantum computers.

Chapter 3

Novel Statistical Methods and Application in Network

3.1 Introduction

Networks have become increasingly popular as representations of complex data because everything is connected now. For example, genes to proteins, proteins to disease, disease to drugs, and drugs to proteins. All these connections are both simple and complex simultaneously. How can we make sense of such data? How do we discover associations between two or three, or hundreds of different entities? What is linked to what else? How do we find inter-dependencies? All those questions can be explored by statistical network analysis. Network Analysis can be defined as 'modeling an entire data set as a network in a graph database to emphasize, reveal, or reflect the relationships or connections (a.k.a edges) between various components or entities (a.k.a nodes). In other words, Network Analysis offers an all-around graphical view of all the links between various nodes. Network Analysis, in the social science, for instance, is the process of investigating social structures through the use of networks and graph theory; in the life sciences, for instance, graphically demonstrates associations between genes, drugs, diseases, and proteins. In all, network analysis is an interactive representation of data analysis used to generate useful insights from results shown in a graphical form. In this chapter, I would use my coauthored article [3] (H. Cheng et al., 2021) to demonstrate some interesting findings on the international non-governmental organizations (hereafter NGOs) network.

[3] some of text in this chapter also appears in my coauthored article

The idea that non-governmental organizations work within networks is central to our understanding of international relations. NGOs gain information

and increase their advocacy power through networking with other concerned actors (Keck & Sikkink, 1998). When organizations are connected, they become more powerful in their struggle against repressive states or polluting governments, among a myriad of other advocacy causes.

Transnational advocacy networks (TANs) are the "most familiar example" of networks in international relations (Hafner-Burton et al., 2009). Made up primarily of non-governmental organizations (NGOs) from around the world, the information, resources, and power of transnational advocacy actors have been argued to increase as a result of their networking behavior (Risse-Kappen et al., 1999). By joining forces around a common cause, NGOs succeed in their collective struggle against repressive states or polluting governments, among the myriad of other advocacy causes.

While much research has shown how networking improves advocacy outcomes, other research has highlighted how the structure of advocacy networks can reinforce global power discrepancies. Even though NGOs are often assumed to increase governance access to the world's powerless, not all NGOs have equal access to or equal involvement with the advocacy network. Organizations from the global South are regularly left out of the overall advocacy network (Fowler, 2000; Shumate & Dewitt, 2008; Townsend & Townsend, 2004). A lack of resources, both human and material, often means that organizations from the global South cannot attend NGO conferences or participate in working groups. When organizations from the global South do participate in the network, many raise concerns about exploitation by their global North counterparts (Kassal et al., 2008; Nelson, 1997). Organizations in the global North may use information on the plight of individuals in the global South for their personal fundraising or mission goals (Bob, 2005; Pallas & Urpelainen, 2013). Moreover, power differentials within the advocacy network may make it difficult for organizations in the global South to have their issues championed by the overall network (Carpenter, 2014).

The troublesome power disparities identified in the structure of the advocacy network are incredibly important in today's political environment. NGOs are currently facing a global backlash and "closing space" in the world community (Brechenmacher, 2017). Repressive regimes have ousted international NGOs for not being well connected to local communities and pursuing their own agenda. Regimes have stopped international aid to NGOs, claiming that this aid makes global South organizations beholden to the foreign policy desires of powerful countries from the global North. The power disparities that seem endemic to the advocacy network could diminish the access NGOs have to cer-

tain affected populations and/or provide a cover through which states could counter the advocacy concerns of civil society actors (Terman, 2019).

Our study seeks to contribute to this growing debate about the utility and structure of the advocacy network for NGOs:

- How power disparities influence the structure of the advocacy network
- How the community detection helps us understand emergent divisions within the advocacy network
- How concepts of brokerage and brokerage roles analysis help us understand how the transnational advocacy network can both extend power to organizations, while at the same time reinforcing power inequalities among NGOs

However, the following chapter is organized as follows. First, we outline the existing literature on transnational advocacy networks, paying special attention to divisions within this literature. Next, we present a cross-disciplinary review of the concepts of community and brokerage and apply these concepts to the study of NGOs. We present our argument and some testable hypotheses that flow from our logic. We then present our novel data and describe our results. Our study concludes with some practical steps that the UN and other concerned actors can take to increase representation and limit divisions within the NGO network.

3.2 Promise and problems of transnational advocacy networks

When compared to the states and corporations that they often seek to change, NGOs are relatively powerless actors in international politics. They do not have standing militaries or deep coffers; many have no paid staff at all. NGOs gather information, frame issues, educate populations, and mobilize global dissent in order to pressure more powerful actors to change behaviors and adopt certain policies, like getting a government to release a political dissident or getting a company to reduce its carbon footprint. As the number of NGOs began to exponentially grow at the end of the Cold War, scholars recognized that organizations do not work in isolation (Brysk, 1993). Successful advocacy often depends on organizations working together with a variety of other actors that share their same advocacy goals. These actors may include local movement leaders, churches, labor unions, parts of various intergovernmental organizations,

and sympathetic government officials. Either internationally or domestically, organizations can also join forces with other NGOs in order to amplify their message and increase their reach.

In their study of international advocacy, Keck and Sikkink (Keck & Sikkink, 1998) call this "dense web of connections" between organizations the transnational advocacy network. In certain situations, the transnational advocacy network increases the advocate's power, leading targeted actors to make concessions in line with the advocate's desires, even if those targeted had previously resisted change. One way that this network can work is through a "boomerang pattern," where connected domestic advocates call out to their transnational network partners to increase international pressure and domestic resources. As DeMars (DeMars, 2005) points out, the overall advocacy network provides additional resources to involved organizations, giving them new tools through which to advocate for change. The transnational advocacy network, with NGOs as the key actor, can be thought of as a public good (Shumate & Dewitt, 2008). It increases the advocacy output and perceived success of involved organizations (Murdie, 2014; Tallberg et al., 2018)

Keck and Sikkink (Keck & Sikkink, 1998)'s characterization of the transnational advocacy network marked a turning point in the study of international politics. Unlike traditional realists, Keck and Sikkink (1998) and later work by Risse, Ropp, and Sikkink (1999), among others, laid out how NGOs and other advocates can become powerful players on the world stage, even without substantial military or material resources. Through this work, even the term "networks" became associated with a structure through which "otherwise weak actors" could "voice their interests and influence governance outcomes in international relations (Avant & Westerwinter, 2016). To note, most of the first wave of work that followed Keck and Sikkink (Keck & Sikkink, 1998) took a "network-as-actor" approach, where "the network is no longer just a way of describing relationships among actors, but an actor until itself" (Kahler, 2011).

The study of the transnational advocacy networks then moved from a "network-as-actor" and to a "network-as-structure" approach, examining relationships within the network (Kahler, 2011; Murdie & Polizzi, 2017). This shift brought many critiques against the somewhat rosy view of NGOs and transnational advocacy networks that had dominated the turn of the millennium. Some of these critiques focused on the assumption that NGOs were more "principled" than other actors in world politics (Cooley & Ron, 2002). Other, more "network-as-structure" critiques centered on how advocacy networks might mirror global power inequalities, with powerful organizations from the global North controlling the advocacy network for their personalistic goals. NGOs in the global

South can become dependent on Northern organizations with more preexisting resources, or may be missing from the overall network in the first place (Jackson, 2020; Murdie, 2014). Savvy organizations and affected populations may have to market their plight to global North organizations interested in their own personal gain (Bob, 2005). Moreover, powerful organizations can act as "gatekeepers," keeping new ideas and issues from permeating through the network (Carpenter, 2014). In addition, the network can include organizations that are free-riders, taking information and resources from others without contributing to the public good (Murdie, 2014). The structure of the network can also affect the strategies taken by organizations, sometimes limiting innovation (Bush, 2015; Hadden & Jasny, 2019; Wong, 2012). To the best of our knowledge, there have been few attempts to reconcile the optimistic view of many of the "network-as-actor" studies with the more pessimistic "network-as-structure" research.

We argue that the network science concepts of community and brokerage provide us with a rich theoretical lens through which to understand the complex nature of the transnational advocacy network. These concepts are relatively new to the growing network literature in IR, especially the literature on NGOs. Below, we first outline how community detection helps us understand divisions within the advocacy network. We then turn our focus to brokerage, a concept which can greatly help us understand how the transnational advocacy network can both extend power to organizations, while at the same time reinforcing power inequalities among NGOs.

3.3 Communities in network

At the most basic level, scholarship on transnational advocacy networks has not envisioned one advocacy network but many separate network structures. Organizations are connected to others in pursuit of common goals; these goals are not uniform across advocacy organizations. As such, an organization should connect to others that share a specific goal or have certain expertise or resources that the organization sees as necessary. Over time, this should create not one "dense web of connection" for NGOs but multiple distinct sub-networks unified over shared values, ideas, or targets (Keck & Sikkink, 1998).

This basic understanding of the creation of transnational advocacy networks molds well with ideas from network science, especially with research on community detection (H.-M. Cheng et al., 2018; Newman, 2006). Networks have emergent properties: as connections occur, new subgroups can emerge within the structure (Maoz, 2017). These endogenous groups or communities

are determined both by the actors themselves but also by the evolving structure of the network.

A community is defined as a "group of nodes that are more tightly connected to each other than they are to the rest of the network" (X. Liu et al., 2019; Mucha et al., 2010). For our purposes, nodes refer to the organizations in the network. Moreover, for our purposes, the network can be thought as the web of connections between organizations generally. Within this overall network, we can identify certain communities that are closely connected. These communities may be driven in part by shared characteristics or issue; however, the communities are also endogenously driven. As Maoz (Maoz, 2017) remarks, these emergent communities form "naturally" over time as a result of the ties that actors develop.

In our empirical models, we rely on a community detection algorithm that will endogenously determine the best composition of communities in our evolving advocacy network over the years (You et al., 2016). Community detection, which is increasingly common in network science, is succinctly summarized by Newman (Newman, 2006):

"Community structure methods normally assume that the network of interest divides naturally into subgroups and the experimenter's job is to find those groups. The number and size of the groups are thus determined by the network itself and not by the experimenter"

Community detection insights do not provide us with testable implications per se, other than the general idea that there will be distinct communities identified in the NGO network. For us, community detection is like data coding: we use community detection methods to identify endogenous communities in our dataset. After using community detection methods, plots, tables, word clouds, and descriptive analysis will help us interpret the legitimacy of the assignment of nodes into various communities. Unlike most existing empirical work on networks of NGOs, we are not assuming that communities are driven only by broad issue-focus or attendance at one specific issue meeting. Community detection methods allow us to remain open to endogenous complexity in community formation. Communities of organizations may be in part driven by advocacy issue or leadership connections (Carpenter, 2014; Henriksen & Seabrooke, 2016), but they also may form for a myriad of reasons particular to the evolving network structure. We think this focus on communities better captures the theoretical insights of the classic "network-as-actor" transnational advocacy network scholarship (Keck & Sikkink, 1998; Risse-Kappen et al., 1999).

3.4 Brokerage in networks

Given that the overall advocacy network is comprised of endogenously emerging communities, how then can we best understand the power disparities identified in the existing "network-as-structure" literature? How can the network both be a powerful actor for international political change and reinforce or exacerbate existing power inequities? We argue that insights from network science and sociological discussions of brokerage are key to understanding this puzzle.

Brokerage is defined as a relationship "involving three actors, two of whom are the actual parties to the transaction and one of whom is the intermediary or broker" (Gould & Fernandez, 1989). We have brokers in many different aspects of our daily lives, from the real estate agent that brokers a deal between buyer and seller to the department head that serves as an intermediary between fussy colleagues. Brokerage is thought to help with innovation (Avant & Westerwinter, 2016). When focusing on advocacy networks, we can think of the broker as the organization that connects two NGOs that otherwise would not have been connected. This could be the regional NGO that takes the interests of a local NGO and frames it for presentation to a big international NGO. It could be the sustainable development organization that participates in discussions both with health and with environmental organizations. Alternatively, it could be the organization that reaches out individually to transitional justice advocates and women's rights advocates after it hears of the abuse of a local individual. This basic idea is at the heart of our understanding of transnational advocacy networks, specifically the boomerang model (Keck & Sikkink, 1998). There are organizations that help in connecting those with needs to those with resources, both internationally and domestically.

Sociologists have highlighted a "dual aspect of brokerage" (Stovel & Shaw, 2012). While brokerage does help transmit information and could help in the pursuit of a specific social, economic, or political goals, brokerage "often breeds exploitation, the pursuit of personal profit, corruption, and the accumulation of power," exacerbating "existing inequalities" (Stovel & Shaw, 2012). Brokers can accumulate power as they control information and access between competitive groups (Burt, 2003). Brokers may control access to communities in ways that stifle innovation. They could selectively relay information for their own goals, ultimately accumulating more power from their brokerage roles. As Stovel and Shaw (Stovel & Shaw, 2012) point out, the idea that brokers benefit from their role is well-established in sociology and network science; brokers can gain "money, information, access to opportunities, enhanced status, or ill-defined claims on side parties' loyalty". This discussion of brokers' accumu-

lation of power is very similar to the "network-as-structure" critiques about transnational advocacy networks: certain organizations enhance their power from the network, acting as gatekeepers or using network information for their gain (Bob, 2005; Carpenter, 2014; L. Jordan & Van Tuijl, 2000).

On the other hand, new research from organizational studies has shown that not all brokers accumulate equal benefits from their brokerage role; certain brokers may lose status in the eyes of their peers. Because brokers are conduits between other actors, they may receive little attention and lack a clear identity themselves, ultimately hurting their status (Sullivan & Stewart, 2017). Crucially, however, the possible negative effect of the brokerage on actor status may depend on the "actor's prior established status" (Sullivan & Stewart, 2017). Actors without a prior high status may see brokerage reduce their status, unlike the general idea that brokerage is a conduit of social power seen in much of the sociological research.

Gould and Fernandez (Gould & Fernandez, 1989)'s influential work serves to connect ideas of community and brokerage. There are different roles that brokers take dependent on whether they are acting as a broker within a specific community or between communities. According to their influential work, in a nondirectional network, there are four potential brokerage roles: coordinator, itinerant broker, gatekeeper/representative, and liaison. Figure 3.1 provides a graphical representation of these relationships.

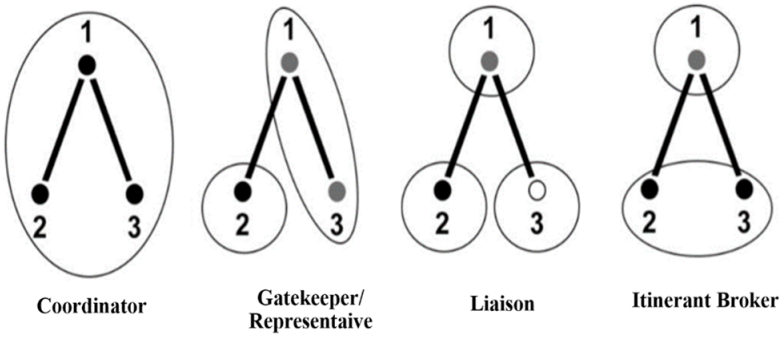

Figure 3.1: Brokerage Roles

A coordinator, also called a local broker, is a broker within one specific community (Gould & Fernandez, 1989). Within its community, it may have more resources or incentives to network than its peers. For NGOs, we could think of this as the organization that knows everyone within a specific area. Its leadership could have a long history working for different NGOs within an area (Henriksen & Seabrooke, 2016). The NGO could transmit ideas or share

strategies between groups that might have trouble connecting directly. The status benefits that this coordinator receives for its role may only be known within the community.

An itinerant broker, also called a cosmopolitan broker, is a broker that is outside a community but is connecting members within that community. An itinerant broker may have special qualities or resources that enable it to connect to two actors within the same community, even when those actors cannot connect themselves. This could be the international NGO that serves to connect two domestic organizations at different sides of a country. This could be the NGO funder that has connections with two competing organizations that rely on outside donations (Cooley & Ron, 2002). Or, this could be an organization that works to represent itself to others outside its area of expertise, like a professional organization (Boli & Thomas, 1999).

A gatekeeper/representative broker connects its community to the outside community. It could be a gatekeeper, being the conduit of information into the community. Or, it could be the representative, being the conduit of information outside the community. Because it controls information and resource flow for members within its community, a gatekeeper/representative broker can accumulate a lot of power and status. It may also be able to manipulate information and resource flows in ways that are personally beneficial. It may develop special skills that help it retain its role. Although not always using these terms, the gatekeeper/representative brokerage role has been talked about extensively in the "network-as-structure" literature on transnational advocacy. This is the NGO that can either facilitate or stop a new issue from making it to the broad advocacy stage (Bob, 2005; Carpenter, 2014).

Finally, a liaison broker connects two actors from separate communities, itself not being a member of either community. Mediators between unions and management would be a classic example of a liaison broker (Fernandez & Gould, 1994). There may be structural reasons why actors from each community cannot or do not connect directly. The liaison broker may have interests that bridge the communities, or it may have resources that allow it to connect when others do not. Liaison brokers may be capable of talking to diverse audiences about a moderate message, like Stroup and Wong (Stroup & Wong, 2017)'s discussion of "leading" international NGOs. According to Stroup and Wong (Stroup & Wong, 2017), leading organizations "receive deference from difference audiences in global politics and therefore have authority". These audiences "can be quite diverse in their preferences," requiring the leading NGO to develop a moderated approach to advocacy (Stroup & Wong, 2017).

We think unpacking brokerage into these distinct brokerage roles can be incredibly useful for international relations. To our knowledge, Gould and Fernandez (Gould & Fernandez, 1989)'s conceptualization of these roles has received limited attention within our subfield.

3.5 Hypotheses

Combining network science and sociological understandings of emergent communities and brokerage roles suggests a much richer picture of the divergent scholarship on transnational advocacy networks. Drawing on this literature, the first goal of our empirical analysis is to identify endogenous communities in the NGO network. Community detection is not a method directly for hypothesis testing (Clauset et al., 2004). Instead, through investigating the results after community detection methods, we can better understand the endogenous process underlying divisions in the advocacy community.

Aided by community detection, we also focus on three testable hypotheses that are implied from our overall theoretical argument connecting communities and brokerage roles in the NGO network. To start, we assume that networking is costly. Organizations with more resources should be more likely to be part of the advocacy network in the first place. Organizations with fewer resources, like comparatively more of the organizations from the global South, may be left out. Although scholarship and organizational initiatives have tried to make it easier for NGOs to network, these power disparities should persist over time.

Within the larger advocacy network, we contend that distinct communities emerge and evolve endogenously, creating separation within the advocacy space. This is consistent with our traditional understanding of advocacy networks: networks are evolving as needs, issues, and actors change (Keck & Sikkink, 1998). These separate communities and the pre-existing power disparities between organizations in the global North and the global South provide an opportunity for some organizations to become more powerful than others as a result of their network position. Although existing work has found that the network may provide resources for all involved (DeMars, 2005; Murdie, 2014; Shumate & Dewitt, 2008), these resources may not be evenly distributed, with some organizations potentially gaining more social power as a result of their particular network position. In general, brokerage positions both provide power and favor the powerful. Although transnational advocacy would never function without brokers, the brokerage also reinforces and exacerbates existing power disparities. This reflects the dual nature of brokerage outlined in Stovel and Shaw (Stovel & Shaw, 2012). Because of this, not only should we see more global North

organizations involved in the network, we should also see more global North organizations as brokers. In short:

Hypothesis 1: Greater participation in the NGO network is associated with global North status.

Second, Gould and Fernandez (Gould & Fernandez, 1989)'s typology of brokerage roles allows us to examine how community and brokerage interact. As communities emerge and evolve, we should see some communities more likely to have organizations acting as particular types of brokers. Some communities may mainly have organizations that are coordinators, rarely venturing out of the specific community. Other communities may be comprised of a disproportionate number of itinerants, gatekeepers/representatives, or liaisons. Over time, evolving communities may develop similar behavioral and interest profiles, leading communities to value certain brokerage roles and have organizations with the social power to retain these brokerage roles. This would both reflect and contribute to growing power disparities within the overall network. This leads to the following testable hypothesis:

Hypothesis 2: There is an association between the community and the distribution of brokerage roles.

And, finally, even within communities with different brokerage role distribution, brokerage roles should be associated with global North status. Becoming an itinerant broker, gatekeeper/representative, or liaison requires resources, both human and material. It requires an organization to be able to speak to divergent audiences, as Stroup and Wong (Stroup & Wong, 2017) point out. Moreover, organizations in these brokerage roles may try to limit the ability of other organizations to take their social power. These ideas suggest that global North organizations may not only dominate the NGO network; they may also dominate certain roles within emerging network communities. This implies:

Hypothesis 3: Within communities, there is an association between global North status and brokerage roles.

The goal of these hypotheses is to provide some basic tests of the implications of our argument. This helps enrich the presentation of the results from the community detection, an approach that is not commonly associated with testable implications per se. We now turn to describe the innovative way we empirically capture NGO networking behavior.

3.6 Data collection

Although theoretically important, NGO networking data has been particularly difficult to gather. Scholars have used many innovative sources and approaches

(Bush, 2015; Carpenter, 2014), and data from the Yearbook of International Organizations (Caniglia, 2001; Murdie, 2014).

We take a somewhat different approach that allows us to examine the NGO network over time. Specifically, we crawled information provided on the UN's "integrated Civil Society Organizations (iCSO) System" during the summer of 2018 for NGO profiles[4].This is a database of over 24, 000 NGOs from all UN member states. Organizations opt into the database when they establish a relationship with the UN's Department of Economic and Social Affairs (DESA) or when they apply for consultative status with the Economic and Social Council (ECOSOC). Participation in UN meetings as an NGO can also lead to an organization having a listing in the database. As DESA's NGO Branch summarizes on its website, in addition to listing an overall organizational objective, NGO profiles include "a general part (name, address, organization type), contacts and meeting participation, activities, and information related to the substantive areas of DESA" (DESA 2019). Figure 2 provides a screenshot of a sample organization's profile. Figure 3.2 provides a screenshot of a sample organization's profile.

[4] https://esango.un.org/civilsociety/login.do

Figure 3.2: Screenshot of iCSO Organizational profile

To obtain NGO-to-NGO network data, we first focused on the meetings that NGOs have attended over time. Figure 3.3 provides a screenshot of meeting participation for a sample organization.

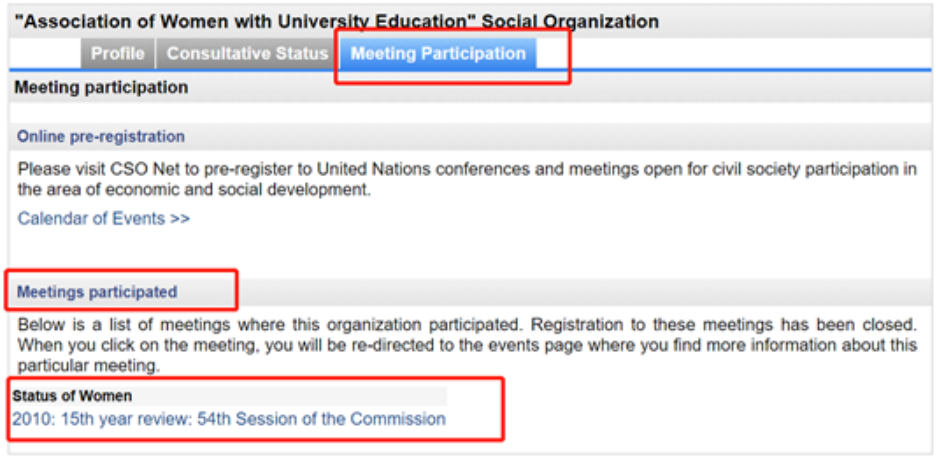

Figure 3.3: Screenshot of iCSO Meeting Participation

During our coding, we were able to identify meetings NGOs attended from 1992 through 2017. We then turned this two-mode (organizations and meetings) network dataset into a one-mode network of organizations connected through joint meeting attendance (Wasserman, Faust, et al., 1994). This is common in community detection (Alzahrani & Horadam, 2016). In total, there were 3, 903 organizations and 1, 300, 519 ties or edges. For our analysis, we deleted edges with a weight of one6 and then isolated nodes were removed, giving us a working dataset of 1, 200 organizations, with edge total varying by year. There were 437, 152 edges identified in the final year of the sample. Figure 3.4 shows how the number of connections between our nodes increases over time. After removing isolated nodes and edges with a weight of one, the remaining NGO organizations rarely registered on the UN's "integrated Civil Society Organizations (iCSO) System" before 2002. As a result, the following analysis focuses on the data from 2002 to 2017.

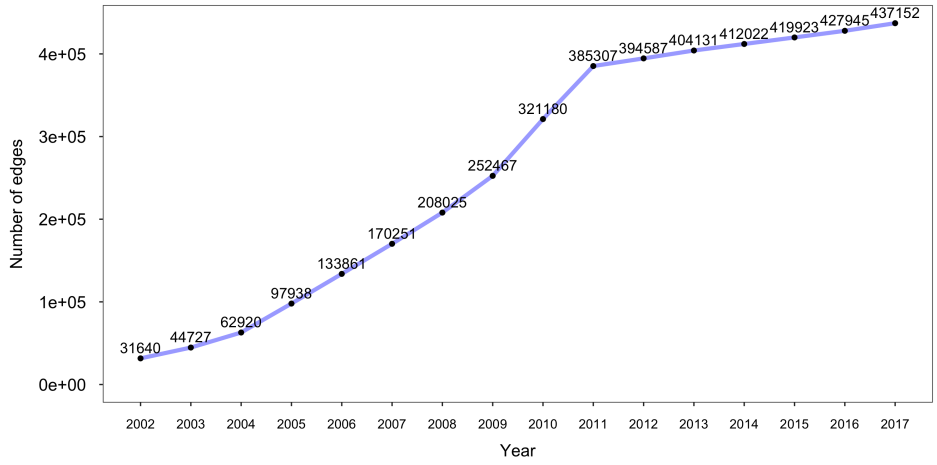

Figure 3.4: The NGO Network over time

We think this dataset will be incredibly useful for future researchers. We also think it matches closely with our central concepts of interest. Joint meeting attendance provides an opportunity for NGOs to strategize, share information, change tactics, address targets, and form partnerships, all practices commonly thought of as NGO networking. Additionally, we think this dataset provides an especially stringent test of our hypotheses. As Moloo (Moloo, 2011) points out, the UN has taken steps over the time period of our sample to increase and ease access for NGOs in the UN system, especially relevant for organizations from the global South. The UN's legitimacy in global governance rests on NGO involvement (Moloo, 2011). As such, the costs of networking for global South organizations in UN meetings may be lower than other comparable networking forums, increasing their likely involvement and biasing our dataset against finding evidence for Hypothesis 1 and 3. Despite this, as discussed below, we find support for our three hypotheses.

3.7 Analysis

3.7.1 Participation in the NGO network

We first extract information on NGO headquarters from the iCSO database and examine whether countries in the global North are overrepresented in the network, as argued in Hypothesis 1. Among the 1200 NGOs in our sample, 1, 116 have addresses listed in iCSO; 102 different countries are listed as the address for these organizations. Figure 3.5 shows the distribution of the top 10 countries of residence of the NGOs in our sample.

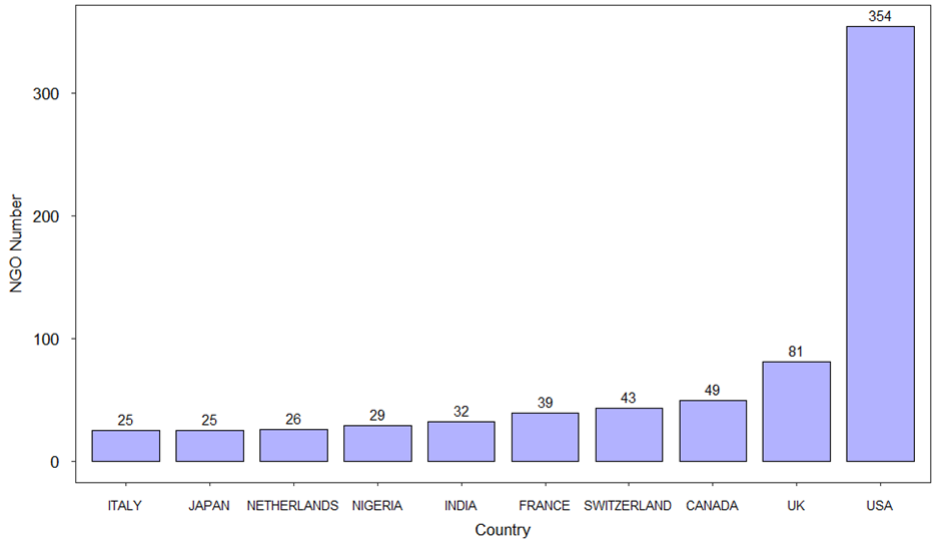

Figure 3.5: Top 10 Country Distribution of NGOs in NGO Network

There is no universal list or definition of countries in the global North. For our study, we code addresses of organizations in the 36 countries that are members of the Organization of Economic Co-operation and Development (OECD) as global North organizations; all other organizations with addresses listed are categorized as organizations from the global South. As Figure 3.6 shows, despite much discussion of the issue of power imbalances in the NGO network, even when focusing just on 2002-2017, after this problem had been widely discussed in the network-as-structure and practitioner literature, we see very little movement in the percentage of global South NGOs involved in the network over time.

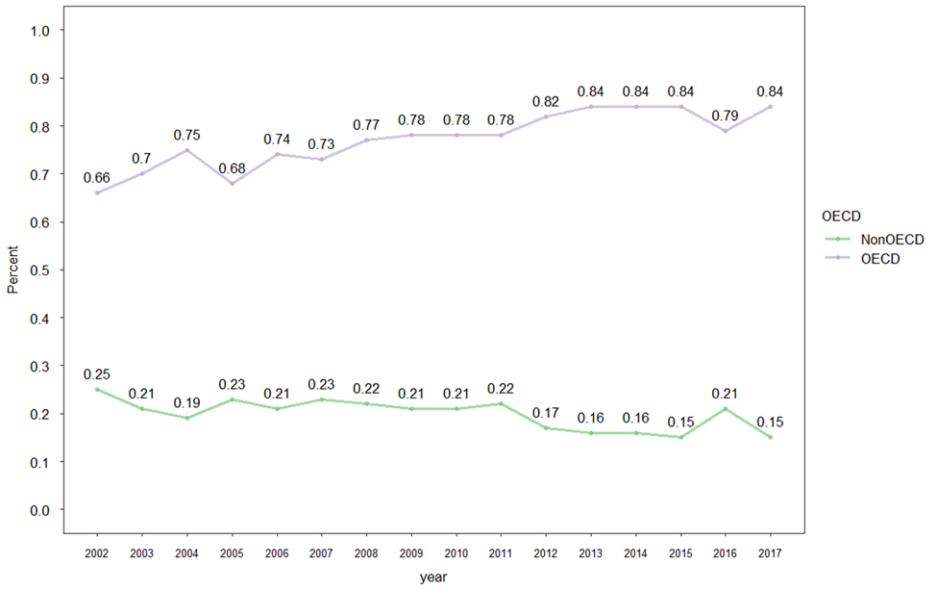

Figure 3.6: Comparing Global South and Global North NGO Percentages in the Network

To test Hypothesis 1, we focus on differences between the average degree centrality scores between organizations from the global North and the global South. Degree is a basic network measure of the total number of ties connecting an actor to others in the total network (Wasserman, Faust, et al., 1994). Figure 3.7 provides an illustration of the distribution of degree over time. Again, to be as stringent as possible, we focus this test only on the years 2002-2017, after there had been widespread discussion of the power disparities between organizations in the global North and global South. Results of Wilcoxon rank sum test ($p < 0.05$) in each year support Hypothesis 1: greater participation in the NGO network is associated with global North status.

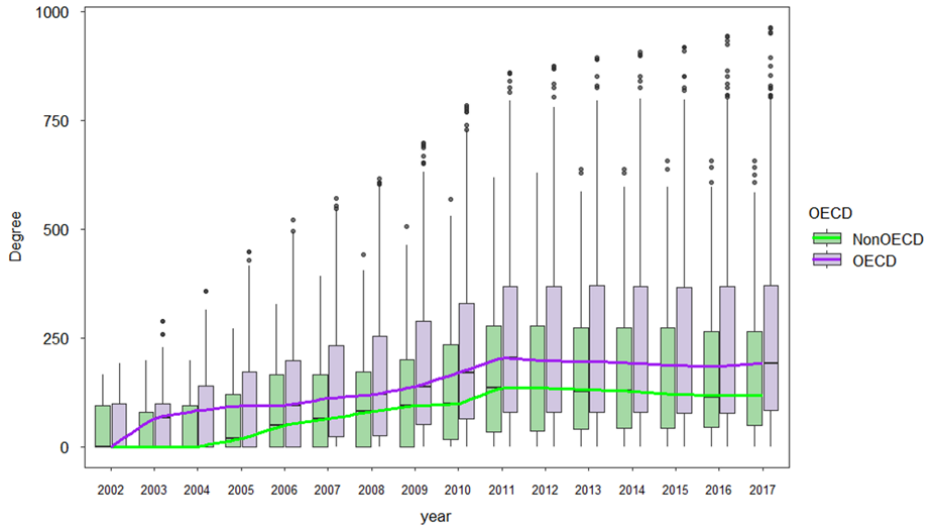

Figure 3.7: Average Degree Centrality over Time within the NGO Network

Another measure of centrality, betweenness centrality, closely relates to ideas of brokerage. In fact, it has been referred to as "brokerage capacity" in extant literature (Caniglia, 2001). Betweenness centrality is thought to capture those that are in the "middle" or "bridges" in that it captures the total number of shortest paths between nodes that go through a particular node (Murdie & Davis, 2012; Wasserman, Faust, et al., 1994). Figure 3.8 provides the OECD and non-OECD distribution of betweenness centrality scores over time. Results of the Wilcoxon rank test sum ($p < 0.05$) in each year also support Hypothesis 1: global North organizations are more likely to participate more in the network, even in ways that resemble brokerage.

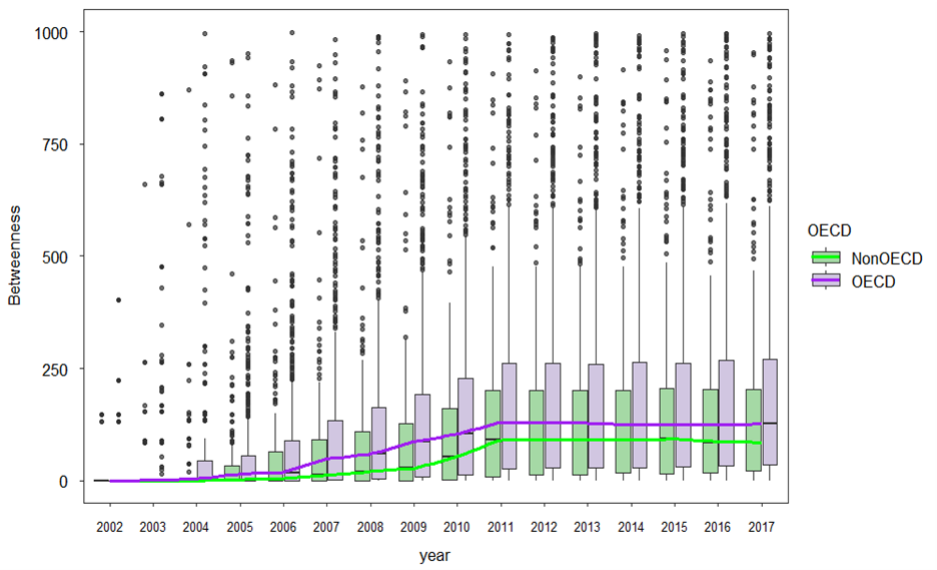

Figure 3.8: Average Betweenness Centrality over Time within the NGO Network

3.8 Community detection

Our next task was to detect emerging communities within the NGO network. We used a time-varying stochastic block model to identify the optimal common composition of communities that maximizes modularity across years. For the static network at time t, the concept of modularity was introduced for networking clustering in 2004 (Newman & Girvan, 2004)

$$modularity(t) = \frac{1}{2m_t} \sum_{\text{i,j in the same community}} [A_{ij}(t) - \frac{k_i(t)k_j(t)}{2m_t}]$$

Where $A_{ij}(t)$ is the number of edges between vertices i and j (the quantities $A_{ij}(t)$ are the elements of the so-called adjacency matrix), $k_i(t)$ and $k_j(t)$ are the degrees of the vertices i and j, m_t is the total number of edges in the network $m_t = \frac{1}{2}\sum_i k_i(t)$, and the leading factor $\frac{1}{2m_t}$ of is merely conventional for normalization. Zhang and Cao (J. Zhang & Cao, 2017) defines the mean modularity of different time as the final modularity of the time-varying network. Modularity varies from 0 to 1. A higher modularity value indicates denser connections between nodes within communities (groups) and sparser connections between communities (H.-M. Cheng et al., 2018; Newman & Girvan, 2004;

J. Zhang & Cao, 2017). In this chapter, we adopt the Louvian maximization method, which is a greedy community detection method (Blondel et al., 2008). The algorithm iteratively builds new communities until the modularity reaches its maximum. Thus, the algorithm automatically provides us a rationale for the number of communities identified.

As shown in Table 3.1, the four identified communities have between 186 and 354 organizations each. Community 3 has the highest mean degree centrality score for its organizations; however, it also has the lowest average betweenness centrality score. Community 4 has the highest mean betweenness centrality. To note, betweenness centrality and degree here record ties to the whole network; our brokerage role analysis, discussed below, discusses how ties differ between and within communities. The modularity value (0.25) of our network suggests many cross-community ties, providing opportunities for brokers.

Table 3.1: Community Detection

Community Label	Community 1	Community 2	Community 3	Community 4
Nodes count	350	310	186	354
Degree Average	143.87	317.32	329.71	167.27
Betweenness Average	550.48	535.89	430.04	628.63

How can we explain the character of these emergent communities? Word clouds of the names of the organizations are shown in Figure 3.9. These help us identify some interesting patterns in the communities. First, Community 1 is composed of many smaller organizations that have names that reflect geographic locations and regions. There appears to be a focus on indigenous or people's rights.

Community 2 is comprised of what many would consider the classic or textbook international NGOs. There are quite a few organizations in Community 2 that are household names: Amnesty International, CARE International, Open Society, and Oxfam, for example. Although there are many mentions of women or women's rights, there are many other organizations that appear generally focused on human rights and development.

Community 3 includes many professional organizations, like the International Sociological Association, the American Psychological Association, and the International Studies Association. It also includes many research-focused

foundations at both ends of the political spectrum, like the Kroc Institute for Peace and Justice, Focus on the Family, the Heritage Foundation, and the International Planned Parenthood Federation. The final community, Community 4, appears to have a large proportion of environmental organizations, including Greenpeace International, Earth Action, Friends of the Earth International, and the Sierra Club.

The composition of the communities detected has a lot of face validity for NGO scholars. Communities 1 and 4 appear to have some issue that may be helping to drive the structure. Community 2 is comprised by many high-status organizations; it is fascinating that the algorithm was able to group so many of these together with no other information other than their network data. Finally, the ideological diversity in Community 3 is especially interesting. In some regards, these organizations are working with very separate and competing agendas.

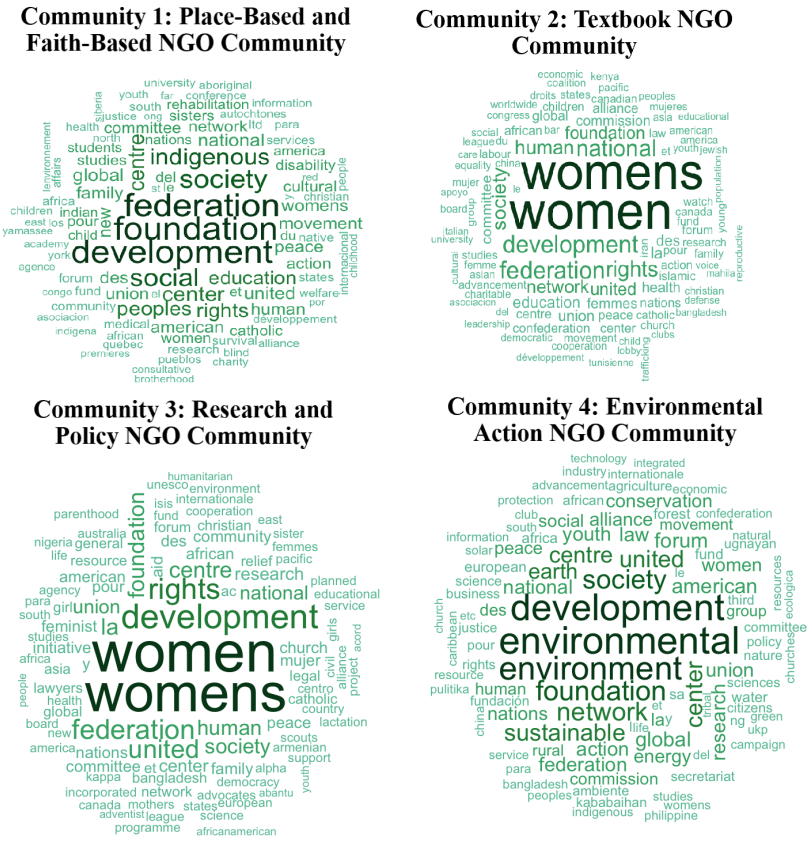

Figure 3.9: Community Grouping

3.9 Brokerage roles

Hypothesis 2 concerns the distribution of brokerage roles across communities. As we argued, some communities are likely to have more coordinators, connecting organizations mainly within a specific community, while other communities could have more itinerants, gatekeepers/representatives, or liaisons, connecting with organizations outside of their specific community in divergent ways. In this way, brokerage provides power to organizations differently across the communities.

Brokers must connect two organizations. As such, only organizations with at least two neighbors were assigned a specific role. Given our communities, brokerage properties were determined in line with Gould and Fernandez (Gould & Fernandez, 1989) using the "brokerage" command in the "sna" R package (Butts et al., 2008). Table 3.2 provides a breakdown of the brokerage role distribution in each community. A Chi-squared test allows us to reject the null hypothesis that there is no association between community and brokerage role at the $p < 0.05$ level, providing support to Hypothesis 2. As expected, there is an association between community and brokerage role.

Table 3.2: Brokerage Role Distribution by Community

Role	Community 1	Community 2	Community 3	Community 4
Coordinator	0.55	0.43	0.00	0.70
Gatekeeper Representative	0.35	0.50	0.49	0.23
Liaison	0.10	0.07	0.29	0.07
Itinerant	0.00	0.00	0.22	0.00

Although the Chi-squared test supported Hypothesis 2 , when taken with a qualitative understanding of the organizations in each community, there are many things in Table 3.2 that are both interesting and surprising. First, Communities 1 and 4 are similar in that they have their largest percentages of brokers in coordinator roles, connecting unconnected organizations within their communities. For both communities, over half of their brokers are coordinators, operating inside the community, instead of outside the community. Community 4, where we identified many environmental organizations, has the highest percentage of brokers (70%) that are coordinators. Perhaps this issue area is

particularly prone to isolation from the overall network; organizations in Community 4 may see limited value in brokering with other communities.

Community 2 was the community with many of the household-name organizations. Although there is still a large percentage of coordinators, indicating that brokers in this community still work to connect others within the community, half of the brokerage roles are gatekeeper/representative roles. This is consistent with much of the "network-as-structure" critiques of these household-name organizations (Bob, 2005; Carpenter, 2014). Somewhat surprisingly, there are no itinerant brokers in this group and a very low percentage of liaison brokers (7%). There are many organizations in Community 2 that Stroup and Wong (Stroup & Wong, 2017) classified as "leading" organizations. However, when compared to Community 3, where there is a higher percentage of liaison and itinerant brokers, Community 2 organizations are somewhat less moderate in their advocacy strategy and may have less need to connect organizations in another community or to connect organizations across communities.

Community 3, where we saw many professional organizations and partisan research organizations, is perhaps the most interesting and divergent in its role distribution. It is the only community to have no coordinators, indicating that brokerage between organizations in this community may be of little value to the organizations. Given the divergent partisan bent of many of the organizations in this community, this could be expected. Community 3 is also the only community where we find itinerant brokers. This could be linked to the large percentage of professional organizations in this community; these organizations often have missions to support the professional interests of their members to the outside world.

Given this discussion of the various distribution of roles across these communities, we now examine the distribution of global North organizations within specific brokerage roles. Within the specific communities, our third hypothesis was that brokerage roles are associated with global North status. Global North organizations may be more likely to get those roles that broker relationships across communities, namely gatekeepers/representatives, liaisons, and itinerant brokers.

It is worth noting that there is no association ($p > 0.05$) between community and OECD status: all communities are made up of between 67% and 74% OECD organizations. In all four communities, however, we do find an association between brokerage role types and OECD status ($p < 0.05$). Figure 3.10 summarizes the percentage of OECD organizations in each brokerage role in each community. The horizontal line in each graph signifies the overall percentage of OECD organizations in each community.

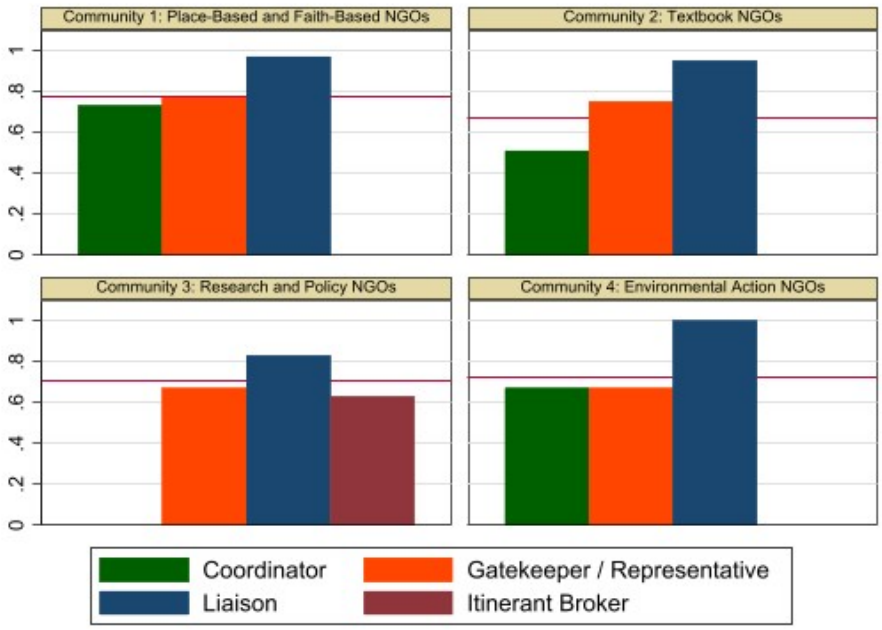

Figure 3.10: OECD Status and Role

In the communities where there are coordinators (Communities 1, 2, and 4), these coordinator roles are filled with the lowest percentages of global North organizations. Gatekeepers and liaison roles are filled with a higher percentage of global North organizations. This is consistent with our expectations and with the extant literature: organizations with pre-existing power have been able to gain additional power through the advocacy network (Bob, 2005). Their brokerage roles could help reinforce their power within their community as a gatekeeper/representative and between multiple communities as a liaison.

As before, Community 3, the community with many partisan and/or professional groups, shows a very different pattern. Since this community is likely to contain organizations whose goals may include reaching outside of professional or issue boundaries, it is somewhat expected that this group has an increased percentage of global South organizations in these brokerage roles. The results concerning itinerant brokers were surprising to us: only 61% of itinerant brokers are from the global North. Perhaps, for organizations in this particular community, networking into a separate community is a special priority, making it the brokerage strategy for both organizations from the global North and the global South. If the organization had a little prior status in Community 3, however, this brokerage role may provide limited status gains in the organization's community (Sullivan & Stewart, 2017).

The social network analysis of this new dataset not only provides support for our hypotheses, it also provides many insights into the evolving structure of networked advocacy and the current NGO environment. The transnational advocacy network may be a public good, increasing perceived success and output (Murdie, 2014; Tallberg et al., 2018). It is also an unequal power environment, with comparatively little participation from the majority of the world's population living outside of the global North.

Further, the network is evolving into distinct communities, with human rights and development organizations that are household names being in one community (Community 2), separate from communities that focus on environmental issues (Community 4), indigenous or regional issues (Community 1), and professional and partisan research organizations (Community 3). Professional and partisan research organizations, although long recognized as part of the INGO community (Boli & Thomas, 1999), have received comparatively little attention in the NGO literature, especially from within international relations. Our examination of community and brokerage, however, shows how different organizations in this community (Community 3) are in their networking behavior from the rest of the communities in our network. This community does not have coordinators and was the only community where we identified itinerant brokers. Further research on this particular community and how it interacts with other NGO communities is definitely necessary.

Our analysis of brokerage roles also provides many insights into the workings of the NGO network and how pre-existing power differences may be associated with certain brokerage roles within communities. In the communities comprised of organizations that have been the traditional focus of international relations (Communities 1, 2, and 4), global North organizations are not only more likely to be involved in the network, they are more likely to take liaison and gatekeeper/representative roles within their community These roles enable organizations to shape the advocacy agenda their community shares with the outside world, ultimately reinforcing and growing their personal power and status. For each community that represents the traditional focus of international relations (Communities 1, 2, and 4), a larger percentage of organizations from the global South are in a coordinator broker role. Although this role may still give these organizations access to the network, a coordinator role will not provide the agenda-setting or resource-allocating power associated with gatekeeper/representative or liaison roles. Through evolving communities and disparities in roles, the advocacy network could be exacerbating power inequalities.

3.10 Chapter contribution

Questions about power are central to international relations and political science; it is critical to understanding who gets what, when, and how (Lasswell, 2018). For a long time, NGOs were considered epiphenomenal to world politics. Later work, however, theorized that NGOs can become powerful conduits of social change through their collective involvement in transnational advocacy networks (Joachim, 2003). Much work has assumed that networks give the powerless a voice (Avant & Westerwinter, 2016). Although networks can help NGOs increase their collective power, our study has shown that the transnational advocacy network may also reinforce power disparities, especially between organizations in the global North and the global South. Distinct communities emerge endogenously. Within these communities, organizations from the global North occupy brokerage roles that help them reinforce their power through resource allocation and information transmission. These findings provide insights into a long-standing puzzle in the NGO literature: namely, why the "network-as-actor" literature sees so much promise in networks while the "network-as-structure" literature sees so much exploitation.

Our study illustrates the utility in incorporating brokerage and community detection into network studies in international relations, especially studies of NGOs. There is a dual nature of brokerage (Stovel & Shaw, 2012). The brokerage allows networks to function, transmitting resources and information to otherwise unconnected actors. However, it also exacerbates power inequalities and can drive manipulation and corruption (Stovel & Shaw, 2012). This dichotomy could be useful to studies of peacekeeping arrangements, rebel, and terrorist groups, or multinational corporations. Further, although existing work within international relations has examined brokerage, to our knowledge, examinations of brokerage have not been divided into separate brokerage roles. Gould and Fernandez (Gould & Fernandez, 1989)'s conceptualization of distinct brokerage roles helps provide leverage on how power, information, and resources may disproportionately flow to some brokers and not others.

Discussions of the community from network science are also informative for NGO researchers. Instead of assuming communities based on issue-focus, community detection algorithms from network science allow communities to emerge endogenously. Our four detected communities have much face validity. Using nothing but information on networking behavior, we were able to identify a community of household name organizations (Community 2) as well as separate communities of environmental (Community 4) and indigenous/regional (Community 1) organizations. We were also able to identify a

community composed of both partisan and professional organizations (Community 3). This community's unique behavior should be the focus of additional research. Community detection could be a useful tool for understanding IGO voting networks and alliance patterns.

Our study offers a cautionary note for NGO practitioners and UN officials interested in ensuring a multitude of civil society voices at UN-facilitated meetings. Despite efforts to the contrary, we find no discernible reduction in the overrepresentation of OECD-based organizations in the NGO network over time. Further, within communities, global North organizations still occupy crucial brokerage roles. These disparities give fuel to repressive regimes interested in limiting civil society space or countering critiques from advocates (Brechenmacher, 2017; Carothers, 2016; Terman, 2019). Efforts to specifically seek global South involvement and representation across communities may be helpful. Grants or donations to foster NGO meeting attendance could also be useful, as long as such grants do not further dependencies and power disparity. Moreover, finally, simply acknowledging that powerful actors still dominate the civil society space and get additional power from networking opportunities could help encourage organizational reflection. To the extent that voices from global South NGOs are critical to UN legitimacy, easing the ability for NGOs to be involved in meetings may help improve public opinion about the UN (Moloo, 2011). Through these changes, the advocacy network may be better equipped to listen to and respond to the plight of the world's powerless.

Chapter 4

Novel Statistical Methods and Application in Biomedicine

4.1 Introduction

Statistical thinking is commonly used in public health and clinical research. In biomedical area, the studies usually involve animals, cell lines and human subjects. Those studies typically include experimental design, sample size calculation, power determination, data collection, data analyses and interpretation, all require statistical support. In this Chapter, I will use my coauthored articles (Phillips et al., 2020; Y. Wang et al., 2020)[5] to illustrate how statistical thinking facilitate the clinical research.

[5] my main jobs include data exploration, data analysis, and data visualization

Obstructive Sleep Apnea (OSA) is the most common cause of sleep apnea and accounts for 75% of all cases of disordered sleep. OSA patients may display abnormally long pauses in breathing or abnormally low levels of breathing during sleep, and often have fragmented sleep, snoring, excessive daytime sleepiness, fatigue, high blood pressure, irritability, depression, loss of concentration, poor neurocognition, and reduced work performance (ARAI et al., 1998; Flemons & Tsai, 1997; Omachi et al., 2009; Rajaratnam et al., 2011; Ward et al., 2009). Because sleep disorders and lack of sleep affect 35% of adult and 68% of adolescent Americans, the CDC has declared sleep deprivation as an epidemic (for Disease Control, Prevention, et al., 2017). Interruptions of breathing, while asleep, result in chronic intermittent low oxygen levels (chronic intermittent hypoxia) and tissue inflammation. Hence, OSA is most commonly treated with Continuous Positive Airway Pressure (CPAP) administered at night while sleeping. Dental airways devices produce a similar treatment effect, and thus, have gained

some recent acceptance as an effective alternative to CPAP (Cantore et al., 2016; Kostrzewa-Janicka et al., 2016).

Hypoxia induced systemic inflammation is often considered the major cause of increased risk for the various apnea-related health problems. These problems develop over time and include cardiovascular disease, metabolic syndrome associated insulin deficiency and diabetes, tissue inflammation, hypertension, obesity, depression, cognitive decline, and stroke, all of which increase mortality (Fava et al., 2011). More recently apnea has been linked to increased risk for a number of autoimmune diseases affecting a variety of tissues and organs including autoimmune encephalitis (Blattner et al., 2019), systemic lupus erythematosus (SLE), rheumatoid arthritis, ankylosing spondylitis, Sjogren's syndrome, and systemic sclerosis, autoimmune hypopituitarism, atopic dermatitis, and psoriasis.

Even though OSA has been linked to the risk of autoimmune disease (B. Abrams, 2005; E Mirrakhimov, 2013; Kang & Lin, 2012), the evidence that chronic intermittent hypoxia-induced chronic inflammation might be the mechanistic cause is only emerging recently (Masarsky, 2018; Serebrovskaya & Xi, 2015; Vakil et al., 2018a). Hypoxia leads to necessary changes in energy metabolism in myeloid cells (Sadiku & Walmsley, 2019) with the potential to influence autoantibody production (Jing et al., 2020). It is well known that in cultured cells hypoxia induces a number of stress signaling cascades mediated by factors including HIF-1, NF-kappa B, and Nrf2, endothelin 1 and VEGF. By one current model, oxidative stress signaling induces higher levels of a number of inflammatory cytokines to initiate an inflammatory cascade, which in turn increases the risk of autoimmune disorders (Vakil et al., 2018b). Inflammatory cytokine cascades (Vakil et al., 2018b) and over stimulation of dendritic cells by autoantigens may stimulate auto-reactive B cells to increase the production of autoantibodies (Toubi & Vadasz, 2019). Altered levels of TNF-Alpha (Andreakos, 2003; Z. Liu et al., 2011; Loftus Jr, 2007), IL-17 (Huang et al., 2016; Kuwabara et al., 2017; Tabarkiewicz et al., 2015) and IL-6 (Ding et al., 2009a; Kishimoto et al., 2015; Tanaka et al., 2014) are all associated with autoimmune disease and appear to play roles in initiating hypoxia-induced inflammatory cascades. Relative to control subjects OSA patients are reported to have 1.2 to 2.5 fold higher levels of, TNF-Alpha, IL-17 and IL-6. Most, but not all, of these studies show airways therapy results in more normal levels of TNF-Alpha, IL-17, and IL-6, more similar to the levels in control individuals without apnea. Based on a model in which these cytokines initiate inflammatory signaling, TNF-Alpha and IL-6 are targets for immunotherapeutic suppression of inflammatory autoimmune

diseases (Andreakos, 2003; Ding et al., 2009b; Kishimoto et al., 2015; Loftus Jr, 2007; Tanaka et al., 2014).

Our goal was to discover other novel autoimmune-associated cytokines that responded to airways treatment in OSA patients and might play roles in autoimmune diseases. Their identification would increase our understanding the link between OSA and autoimmunity and these cytokines might be additional targets for therapeutic treatment of OSA. We compared the levels of several inflammatory cytokines previously linked to autoimmunity in serum among well matched patients with OSA not yet receiving airway therapy to OSA patients receiving airways therapy, and also to healthy control individuals. We found significant changes in the levels of APRIL (*TFNSF13*), CD30 (*TNFRSF8*), IFN-Alpha-2 (*IFNA2*), and IL-2 (*IL2*) in OSA patients receiving airway therapy, such that cytokines were more similar to the levels observed in healthy control subjects.

Moreover, the following chapter is organized as follows. First, some details about patient data are explored. Next, I apply statistical test to identified four cytokines associated with autoimmune disease, whose median serum levels were significantly different for OSA patients receiving airways therapy, from the levels in untreated OSA patients, APRIL (5.2-fold lower, $p = 3.5 \times 10^{-11}$), CD30 (1.6-fold higher, $p = 7.7 \times 10^{-5}$), IFN-Alpha-2 (2.9-fold higher, $p = 9.6 \times 10^{-14}$) and IL-2 (1.9-fold higher, $p = 3 \times 10^{-4}$). Then I present that t-SNE and UMAP analysis of these high dimensional patient cytokine data identified only two groups, suggesting a similar global response for all four cytokines to airways therapy. These findings suggest the levels of these four cytokines may be altered by disordered sleep and perhaps by chronic hypoxia.

4.2 Data exploration

4.2.1 Patient data

Nineteen OSA patients had formerly been diagnosed using polysomnography (PSG) based on their Apnea Hypopnea Index (AHI > 5), but were currently receiving nightly airways therapy and were designated airways treated OSA patients. Eighteen of these recorded using CPAP, while patient 31 reported using a dental airways device (Cantore et al., 2016). There were 19 OSA patients currently with apnea, but not receiving airways therapy. Also 8 Control individuals were recruited, but among these, patient 9 was borderline for high blood pressure (PB $145/89$). Although autoimmune disorders can develop in younger or older individuals, none of our patients reported having an autoimmune dis-

ease. Neither patient 31 nor patient 9 produced outlying data. Patients were evaluated in this study after obtaining written informed consent. Gender, BMI, CVD, age, ESS (M. Johns & Hocking, 1997; M. W. Johns, 1991), AHI, SaO2% low, glucose levels, Cholesterol, LDL, HDL, and CRP, were assessed at the time of first recruitment in Table 4.1, which in the case of the airways treated patients was after months of treatment. A Yes/No indication was recorded for chronic medications. Compliance with nightly airways therapy was confirmed by patients' response to a simple yes/no question. The airways therapy treated OSA patients and untreated OSA patients were well matched for nearly all parameters. The control group was considerably younger and leaner, and potentially represented more nearly optimal biometric data and cytokine levels. Patients were recruited, consented, and blood drawn at the University of Georgia's Clinical and Translational Research Unit (CTRU) in Athens, GA.

Table 4.1: Summary of patient biometric, sleep and laboratory data

	Control subjects (n=8)	Airways treated patients (n=19)	Apneic patients (n=19)
Female/Male	6/2	7/12	7/12
Age	37.7 ± 11.5	60.6 ± 10.5	58.2 ± 12.4
Hypertension or heart disease Y/N	2 Yes / 6 No	11 Yes/8 No	11 Yes/8 No
Race C/H/B/A	4C/0H/3B/1A	17C/0H/2B/0A	11C/1H/6B/1A
BMI	26.8 ± 5.96	33.1 ± 9.07	35.0 ± 9.22
AHI at time of diagnosis	1.58 ± 1.64	35.7 ± 24.8	26.8 ± 25.7
SaO2 low %	$91.5\% \pm 2.8\%$	$80.7\% \pm 5.4\%$	$76.4\% \pm 10.3\%$
ESS	5.33 ± 3.67	7.16 ± 5.80	7.94 ± 4.02
Glucose mg/dL	94.8 ± 8.70	106 ± 18.7	104 ± 12.1
Cholesterol mg/dL	165 ± 26.8	181 ± 25.1	179 ± 43.1
HDL mg/dL	52.1 ± 14.6	51.8 ± 18.1	45.7 ± 17.3
LDL mg/dL	96.4 ± 22.6	104 ± 27.6	106 ± 35.8
hs-CRP mg/L	1.07 ± 0.689	4.84 ± 7.65	3.48 ± 5.60
Chronic Meds Y/N	2 Yes/6 No	16 Yes/3 No	16 Yes/3 No
Airways therapy adherence	0 Yes/8 No	19 Yes/0 No	0 Yes/19 No

4.2.2 Cytokine levels

The levels of inflammatory cytokines were examined using Bio-Plex^{TM} Pro Human Inflammation Panel 1 multiplex kits that quantify biomarkers of human inflammation (BioRad 171AL001M4). Multiple 96 well plates were assayed using Bio-Rad Bio-Plex instrument at UGA's Cytometry Shared Resource Laboratory. The Bio-Plex system has the advantage that hundreds of individual beads each estimate each cytokine level in each well, which improves the statistical accuracy of each individual well estimate of all cytokines assayed. All the flash frozen serum samples were thawed only once. Serum, standards and assay controls were diluted as per the manufacturer's instructions (Bio-Rad Bulletin 10,044,281). Table 4.1. As recommended, each serum sample was diluted 4-fold. Fifty microliters of this dilution were run in triplicate for the 8 control, 19 untreated OSA patients, and 19 airways therapy treated OSA patients, instead of running duplicate patient serum samples recommended by the manufacturer. This allowed a more robust assessment of potential experimental errors in each cytokine assayed. The picogram output data for each serum cytokine level was normalized to the concentration of standards, run as an eight-step, four-fold dilution series of each cytokine, and run in duplicate on each assay plate. The quantitative nature of these assays over the expected concentration ranges estimated for serum samples was confirmed by comparing the fluorescence output of the quadruplicate standard samples (two from each plate). The standard error of the lowest concentration standards used to estimate concentration was less than 15% and less than that for higher concentrations.

4.3 Statistical analysis

The data from separate plates were combined to make multiple excel data files, one for each cytokine. The levels of IFN-Alpha-2 among some untreated OSA patients and five airways treated patients were either at or were below the range of detection, with the latter being designated as out of range, $OOR <$, by the instrument software. The lowest picogram patient serum sample concentration of IFN-Alpha-2 that was estimated to be in the range of detection was substituted for $OOR <$ cytokine values. In this way, any estimate of fold difference for IFN-Alpha-2 between OSA patients and airways therapy treated OSA patients would not over-estimate the actual fold differences. At this point, the data were moved into R v3.5.1 for further statistical analysis. The data for patient groups were visualized using Boxplot.

After applying the Kolmogorov–Smirnov test in R (Marsaglia et al., 2003; Oztuna, n.d.; N. Smirnov, 1948; N. V. Smirnov, 1939; Zar, 1972), it was clear that the airways treated patient data for cytokine levels were not normally distributed ($p < 0.05$) and often fell into two groups of values. The Kolmogorov–Smirnov test is a nonparametric goodness-of-fit test and could be used to determine whether an underlying probability distribution differs from the hypothesized normal distribution. Therefore, without the normality assumption, the nonparametric Wilcoxon rank-sum test was used to estimate p values for the significance of pairwise differences in cytokine levels among OSA patients, airways treated OSA patients, and controls. The two-sample Wilcoxon rank sum test is a rank-based test that compares values for two groups. Without any distribution assumption, the test addresses if it is likely that an observation in one group is different from an observation in the other, with significance level $\alpha = 5\%$ in our case (Vargha & Delaney, 2000).

To visualize the high-dimensional data for the levels of all four cytokines among all patients and controls in a two-dimensional map, the nonparametric t-distributed Stochastic Neighbor Embedding (t-SNE) visualization method (Maaten & Hinton, 2008) was applied using the Rtsne version 0.15 statistical R package available online (Rtsne, 2017). T-SNE reduces dimensionality by first using a Gaussian distance to analyze the similarity among data points in high-dimensional space and then projecting these data into two dimensional space (Dorrity et al., 2020). We also employed Uniform Manifold Approximation and Projection (UMAP) as an alternative method to visualize these high-dimensional data in two-dimensions (Becht et al., 2019; Dorrity et al., 2020). UMAP analysis is quite distinct from t-SNE in that it first estimates a topology for the high-dimensional data and then uses the topology information to construct two dimensional space. It has been argued that UMAP may be superior to and/or equivalent to t-SNE at recovering the global structure among high dimensional data (Kobak & Linderman, 2019).

4.4 Results

Cytokine levels were examined in the serum of nineteen OSA patients receiving airways therapy (airways treated patients, Table 4.1) and compared to nineteen OSA patients not receiving airways therapy and a group of volunteers without OSA (control individuals). Table 4.1 summarizes important biometric, sleep and laboratory data for the three groups of subjects with details.

We found the levels of four cytokines with previously reported rolls in autoimmunity were significantly different in airways treated OSA patients relative

to untreated OSA patients, including APRIL (*TFNSF13*), CD30 (*TNFRSF8*), IL-2 (*IL2*) and IFN-Alpha-2 (*IFNA2*) with the detailed data presented as four box plots in Fig 4.1. Increased serum levels of APRIL (Tumor Necrosis Factor Superfamily Member 13) and increased APRIL signaling have been linked to several autoimmune disorders including rheumatoid arthritis (Boghdadi et al., 2015), eczema (Mohamed Ezzat et al., 2016), multiple sclerosis (Thangarajh et al., 2005), and systemic lupus erythematosus (Samy et al., 2017). We found the median pg/mL serum level of the soluble isoform of APRIL was 5.2-fold lower in airway treated OSA patients relatively to the median level in untreated OSA patients ($p = 3.5 \times 10^{-11}$, Fig 4.1 A). The level in airways treated patients was still 1.7-fold higher than the levels in control subjects ($p = 6.6 \times 10^{-8}$). Hence, patients receiving airways therapy had levels of APRIL that were much closer to those in controls, but still higher than the levels observed in controls.

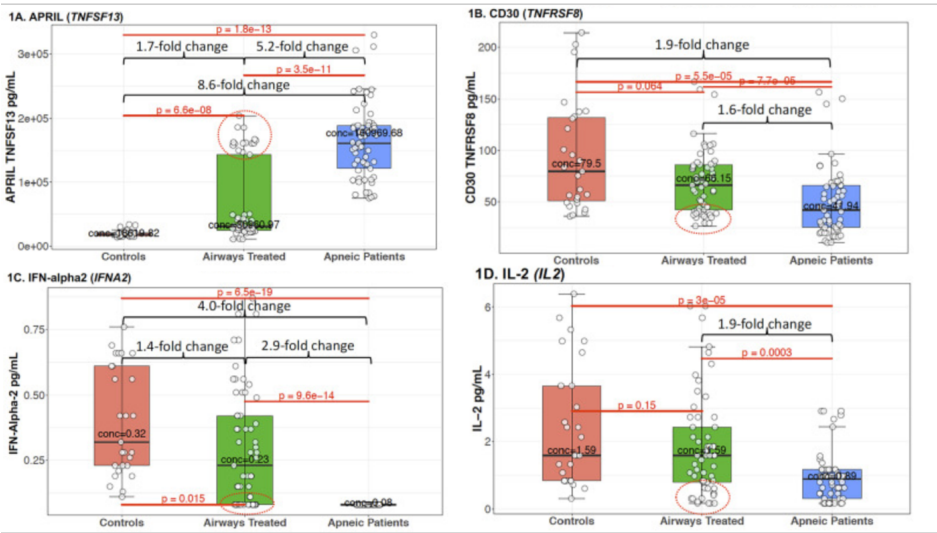

Figure 4.1: The levels of four cytokines involved in the autoimmune disease are significantly altered in OSA patients receiving airways therapy. The serum picogram per milliliter (pg/mL) levels of (A) APRIL, (B) CD30, (C) IFN-Alpha-2 and (D) IL-2 from control individuals, airways treated OSA patients and OSA patients not receiving airways therapy are summarized in box blots. The top box encloses the third quartile and is bounded by median pg/mL value, the lower box encloses first quartile and is bounded by median value. The whiskers indicate the median values $\pm$ 1.5 IQR (interquartile range), and hence exclude outliers. The median value is indicated by a black line. Each of the three independent estimates of a cytokine level for each patient are represented by separate data points. Outliers among the airways treated patients that resemble untreated OSA patient data are encircled with a red dotted line.

CD30 (CD30L) is member of the Tumor Necrosis Factor Receptor Superfamily expressed in activated T and B cells. Its soluble isoform is generally found to be upregulated in leukocytes in patients with chronic inflammatory and autoimmune diseases including lupus erythematosus, asthma, rheumatoid arthritis and atopic dermatitis and CD4+ T cell-mediated graft-versus-host disease (Blazar et al., 2004; Oflazoglu et al., 2009). However, we found the median pg/mL level of CD30 was 1.6-fold higher in the serum of airways treated OSA patients relative to untreated OSA patients ($p = 7.7 \times 10^{-5}$, Fig 4.1 B). Whereas the median level in airways treated patients was slightly lower than that in controls, this difference was not statistically significant ($p = 0.064$).

Interferon Alpha 2 (IFN-Alpha2) expression is elevated in a number of autoimmune diseases such as arthritis, systemic lupus erythematous and Sjogren's syndrome with the proposed effect of reducing both inflammation and the autoimmune response (Hall & Rosen, 2010; Kim & Moudgil, 2017). The median pg/mL level of the soluble isoform of INF-alpha2 was 2.9 -fold higher in the serum of airways treated OSA patients relative to untreated OSA patients ($p = 9.6 \times 10^{-11}$, Fig 4.1 C). Median IFN-Alpha-2 levels in airways treated OSA patients were more similar to the levels in control subjects, but were statistically distinguishable ($p = 0.015$)

Defects in T Cell Growth Factor Interleukin IL-2 (IL2) or in IL-2 signaling produce multiorgan autoimmunity and are linked to systemic lupus erythematosus (Humrich & Riemekasten, 2016), asthma (Movahedi et al., 2008), and multiple sclerosis (Cannella & Raine, 1995; Shokrgozar et al., 2009). We found the median pg/mL serum level of IL-2 was 1.9-fold higher for airways treated OSA patients relatively to untreated OSA patients ($p = 0.0003$, Fig 4.1 D). The median level in airways treated OSA patients was not statistically distinguishable from the median level in controls subjects ($p = 0.15$).

In short, it appears that OSA patients had aberrantly low levels of all four autoimmune-related cytokines, and OSA patients receiving airways therapy had cytokine levels more similar to those observed in control subjects. However, the cytokine levels for airways treated OSA patients did not appear normally distributed and the data for five patients appeared as outliers with levels that were more similar to those of untreated patients (red encircled data, Fig 4.1). In order to visualize the potentially coordinated response of all four cytokine levels independent of the direction of that response for most of airways treated OSA patients relative to OSA patients and controls and the potential common relationship among the outliers, we applied two machine-learning 2D visualization strategies (Fig 4.2). The first, t-SNE is a non-linear dimensionality reduction method, that has the capacity to capture local structures among these high-

dimensional data, while also revealing the presence of groups of related data as global, and to present these data in two-dimensional space (Maaten & Hinton, 2008).

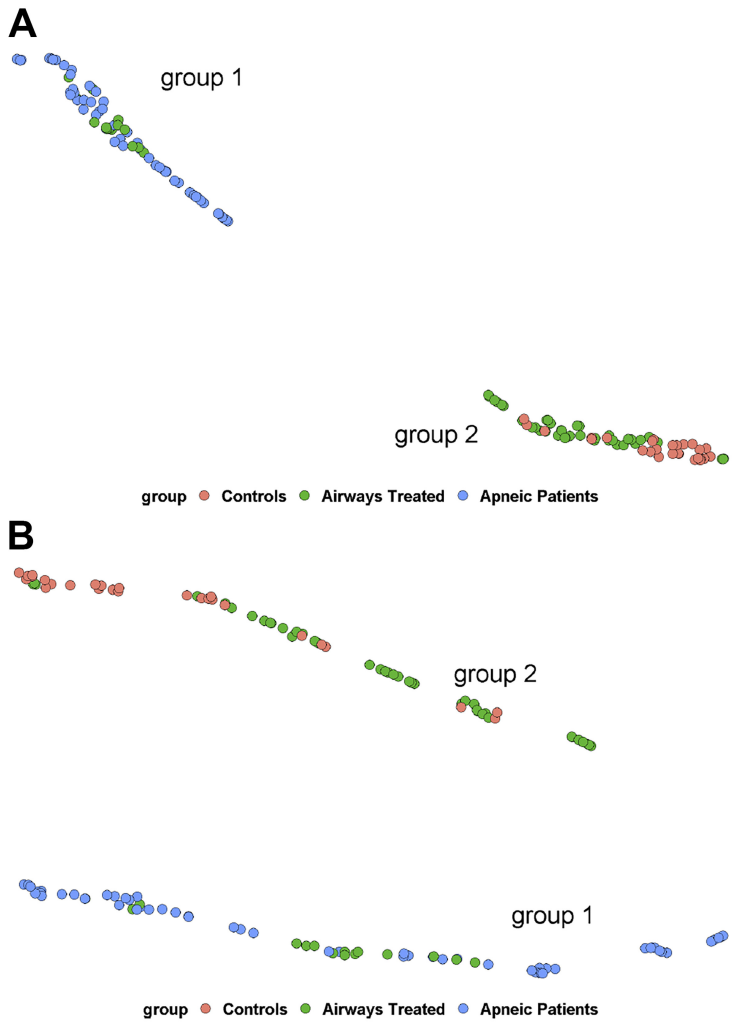

Figure 4.2: t-SNE and UMAP placed the high dimensional patient cytokine data into the same two groups in two dimensions. The high dimensional data for the changes in the levels of APRIL, CD30, IFN-Alpha-2 and IL-2 were reduced to a two dimensional visualization by t-SNE (A) and UMAP (B). A. t-SNE. Group 1 represents the dimensional distribution of the levels of four cytokines among all the untreated OSA patient and a five of the nineteen patients airways treated OSA patients. Group 2 represents the cytokine data for all the control individuals and fourteen of the nineteen airways treated OSA patients. B. UMPA. Group 1 and 2 in have the same affiliated patients as observed with t-SNE. Each patient is represented by 34 data points that combine the dimensional distribution of the levels of the three measurements made for each of four cytokines for that patient.

As shown in the t-SNE analysis in Fig 4.2 A, all the patient cytokine data lie in two clusters, with Group 1 representing the cytokine data among all the untreated OSA patients and five of the nineteen airways treated OSA patients. Group 2 represents the cytokine data for all the controls and fourteen of the nineteen airways treated OSA patients. The robustness of the t-SNE result was tested by applying UMAP, an alternative method for recovering global structure among high dimensional data. Applying UMAP to these patient data (Fig 4.2) we again found two groups. Group membership is the same for Group 1 and Group 2 data using either t-SNE or UMAP.

4.5 Chapter conclusion

Sleep apnea is highly associated with increased risk of various autoimmune diseases. Three cytokines, TNF-Alpha, IL-17, and IL-6, that are positively associated with autoimmune disorders are often elevated in OSA patients and decreased in response to airways therapy. Herein, and fitting this pattern, we find the levels of APRIL were relatively high in apneic patients, but were significantly reduced by airways therapy. Airways therapy did not reduce APRIL all the way to the very low levels observed in controls. By contrast, the levels of CD30, IFN-Alpha-2, and IL-2 were relatively low in OSA patients and were significantly increased by airway therapy, increased to levels similar or statistically indistinguishable from controls. Using t-SNE to analyze our high-dimensional data for the levels of all four cytokines among all subjects, we found that most airways treated patients clustered with the control individuals, while data from the five outlying airways treated patients clustered with the apneic patients. The clustering of these high-dimensional patient data for all four autoimmune-associated cytokines suggests their miss-expression in OSA patients may be linked to increased risk of autoimmune disease.

The low levels of CD30, IL-2 and IFN-Alpha-2 we observed in OSA patients contrasted with expectations of increase in their expression based on previous direct or indirect evidence linking their elevated expression with acute hypoxia. APRIL levels were higher in OSA patients than in airways treated OSA patients, but the link between APRIL expression and hypoxia experienced by OSA has not been suggested in previous literature. Perhaps the chronic intermittent hypoxia experienced for months and years by OSA patients leads to an attenuation of the acute response and chronically altered levels of all four cytokines. This distinction, that cytokine levels respond differently to chronic intermittent hypoxia experienced by OSA subject than to acute hypoxia in experimental systems was described previously (Vakil et al., 2018a). The fact that all four

were more similar to control levels in airways treated OSA patients suggested the likely link to blood oxygenation levels. However, the lack of a correlation between cytokine levels and either SaO_2 low % or CRP levels of any patient group or groups draws into question any clear conclusion about the direct role of hypoxia. Airways therapy of OSA patients appears to be an effective way to control aberrant levels of these four cytokines involved in autoimmune disease and immune processes. The outlying cytokine data for five airways treated patients, suggests we may need a more critical method to access compliance with airways therapy.

Bibliography

Abrams, B. (2005). Long-term sleep apnea as a pathogenic factor for cell-mediated autoimmune disease. *Medical hypotheses*, *65*(6), 1024–1027.

Abrams, D. S., & Lloyd, S. (1997). Simulation of many-body fermi systems on a universal quantum computer. *Physical Review Letters*, *79*(13), 2586.

Alzahrani, T., & Horadam, K. J. (2016). Community detection in bipartite networks: Algorithms and case studies. *Complex systems and networks* (pp. 25–50). Springer.

Andreakos, E. (2003). Targeting cytokines in autoimmunity: New approaches, new promise. *Expert opinion on biological therapy*, *3*(3), 435–447.

ARAI, H., FURUTA, H., KOSAKA, K., KANEDA, R., KOSHINO, Y., Sano, J., KUMAGAI, S., & YAMAMOTO, E. (1998). Changes in work performances in obstructive sleep apnea patients after dental appliance therapy. *Psychiatry and clinical neurosciences*, *52*(2), 224–225.

Avant, D., & Westerwinter, O. (2016). *The new power politics: Networks and transnational security governance*. Oxford University Press.

Beale, E., Kendall, M., & Mann, D. (1967). The discarding of variables in multivariate analysis. *Biometrika*, *54*(3-4), 357–366.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, *37*(1), 38–44.

Bennett, C. H., Bernstein, E., Brassard, G., & Vazirani, U. (1997). Strengths and weaknesses of quantum computing. *SIAM journal on Computing*, *26*(5), 1510–1523.

Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, *44*(2), 813–852.

Blattner, M. S., de Bruin, G. S., Bucelli, R. C., & Day, G. S. (2019). Sleep disturbances are common in patients with autoimmune encephalitis. *Journal of neurology*, *266*(4), 1007–1015.

Blazar, B. R., Levy, R. B., Mak, T. W., Panoskaltsis-Mortari, A., Muta, H., Jones, M., Roskos, M., Serody, J. S., Yagita, H., Podack, E. R., et al. (2004).

Cd30/cd30 ligand (cd153) interaction regulates cd4+ t cell-mediated graft-versus-host disease. *The Journal of Immunology*, *173*(5), 2933–2941.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Blumensath, T., Yaghoobi, M., & Davies, M. E. (2007). Iterative hard thresholding and l_0 regularisation. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, *3*, III–877.

Bob, C. (2005). *The marketing of rebellion: Insurgents, media, and international activism*. Cambridge University Press.

Boghdadi, G., El-Sokkary, R. H., Elewa, E. A., & Abbas, S. F. (2015). April level as a marker of disease activity in treated rheumatoid arthritis patients: Association with disease activity and anti-ccp antibody. *Egypt J Immunol*, *22*, 31–39.

Boli, J., & Thomas, G. M. (1999). *Constructing world culture: International nongovernmental organizations since 1875*. Stanford University Press.

Boyer, M., Brassard, G., Høyer, P., & Tapp, A. (1998). Tight bounds on quantum searching. *Fortschritte der Physik: Progress of Physics*, *46*(4-5), 493–505.

Brechenmacher, S. (2017). *Civil society under assault: Repression and responses in russia, egypt, and ethiopia* (Vol. 18). Carnegie Endowment for International Peace Washington, DC.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, *24*(6), 2350–2383.

Brysk, A. (1993). From above and below: Social movements, the international system, and human rights in argentina. *Comparative Political Studies*, *26*(3), 259–285.

Burt, R. S. (2003). The social structure of competition. *Networks in the knowledge economy*, *13*, 57–91.

Bush, S. S. (2015). *The taming of democracy assistance*. Cambridge University Press.

Butts, C. T. et al. (2008). Social network analysis with sna. *Journal of statistical software*, *24*(6), 1–51.

Byrnes, T., & Yamamoto, Y. (2006). Simulating lattice gauge theories on a quantum computer. *Physical Review A*, *73*(2), 022328.

Candes, E., & Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, *35*(6), 2313–2351.

Caniglia, B. (2001). Informal alliances vs. institutional ties: The effects of elite alliances on environmental tsmo networks. *Mobilization: An International Quarterly*, *6*(1), 37–54.

Cannella, B., & Raine, C. S. (1995). The adhesion molecule and cytokine profile of multiple sclerosis lesions. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *37*(4), 424–435.

Cantore, S., Ballini, A., Farronato, D., Malcangi, G., Dipalma, G., Assandri, F., Garagiola, U., Inchingolo, F., De Vito, D., & Cirulli, N. (2016). Evaluation of an oral appliance in patients with mild to moderate obstructive sleep apnea syndrome intolerant to continuous positive airway pressure use: Preliminary results. *International journal of immunopathology and pharmacology*, *29*(2), 267–273.

Carothers, T. (2016). Closing space for international democracy and human rights support. *Journal of Human Rights Practice*, *8*(3), 358–377.

Carpenter, C. (2014). *"lost" causes: Agenda vetting in global issue networks and the shaping of human security*. Cornell University Press. http://www.jstor.org/stable/10.7591/j.ctt5hhor5

Chen, S., & Donoho, D. (1994). Basis pursuit. *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, *1*, 41–44.

Cheng, H.-M., Ning, Y.-Z., Yin, Z., Yan, C., Liu, X., & Zhang, Z.-Y. (2018). Community detection in complex networks using link prediction. *Modern Physics Letters B*, *32*(01), 1850004.

Cheng, H., Wang, Y., Ma, P., & Murdie, A. (2021). Communities and Brokers: How the Transnational Advocacy Network Simultaneously Provides Social Power and Exacerbates Global Inequalities [sqab037]. *International Studies Quarterly*. https://doi.org/10.1093/isq/sqab037

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

Cooley, A., & Ron, J. (2002). The ngo scramble: Organizational insecurity and the political economy of transnational action. *International security*, *27*(1), 5–39.

DeMars, W. E. (2005). *Ngos and transnational networks: Wild cards in world politics*. Pluto Press London.

Ding, C., Cicuttini, F., Li, J., & Jones, G. (2009a). Targeting il-6 in the treatment of inflammatory and autoimmune diseases. *Expert opinion on investigational drugs*, *18*(10), 1457–1466.

Ding, C., Cicuttini, F., Li, J., & Jones, G. (2009b). Targeting il-6 in the treatment of inflammatory and autoimmune diseases. *Expert opinion on investigational drugs*, *18*(10), 1457–1466.

Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S., & Trapnell, C. (2020). Dimensionality reduction by umap to visualize physical and genetic interactions. *Nature communications*, *11*(1), 1–6.

Draper, N., & Smith, H. (1966). Applied regression analysis. j. wiley & sons inc., ny.

E Mirrakhimov, A. (2013). Obstructive sleep apnea and autoimmune rheumatic disease: Is there any link? *Inflammation & Allergy-Drug Targets (Formerly Current Drug Targets-Inflammation & Allergy)*, *12*(5), 362–367.

Efroymson, M. (1960). Multiple regression analysis. *Mathematical methods for digital computers*, 191–203.

Fan, J., Guo, Y., & Zhu, Z. (2020). When is best subset selection the "best"?

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., Xue, L., & Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, *42*(3), 819.

Fava, C., Montagnana, M., Favaloro, E. J., Guidi, G. C., & Lippi, G. (2011). Obstructive sleep apnea syndrome and cardiovascular diseases. *Seminars in thrombosis and hemostasis*, *37*(03), 280–297.

Fernandez, R. M., & Gould, R. V. (1994). A dilemma of state power: Brokerage and influence in the national health policy domain. *American journal of Sociology*, *99*(6), 1455–1491.

Flegal, K. M., Shepherd, J. A., Looker, A. C., Graubard, B. I., Borrud, L. G., Ogden, C. L., Harris, T. B., Everhart, J. E., & Schenker, N. (2009). Comparisons of percentage body fat, body mass index, waist circumference, and waist-stature ratio in adults. *The American journal of clinical nutrition*, *89*(2), 500–508.

Flemons, W. W., & Tsai, W. (1997). Quality of life consequences of sleep-disordered breathing. *Journal of allergy and clinical immunology*, *99*(2), S750–S756.

for Disease Control, C., Prevention et al. (2017). Sleep and sleep disorders: Data and statistics. *Reviewed May*, *2*.

Fowler, A. (2000). Ngo futures: Beyond aid: Ngdo values and the fourth position. *Third World Quarterly*, *21*(4), 589–603. http://www.jstor.org/stable/3993366

Gould, R. V., & Fernandez, R. M. (1989). Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological methodology*, 89–126.

Grover, L. K. (1997). Quantum mechanics helps in searching for a needle in a haystack. *Physical Review Letters*, *79*(2), 325.

Hadden, J., & Jasny, L. (2019). The power of peers: How transnational advocacy networks shape ngo strategies on climate change. *British Journal of Political Science*, *49*(2), 637–659.

Hafner-Burton, E. M., Kahler, M., & Montgomery, A. H. (2009). Network analysis for international relations. *International organization*, *63*(3), 559–592.

Hall, J. C., & Rosen, A. (2010). Type i interferons: Crucial participants in disease amplification in autoimmunity. *Nature Reviews Rheumatology*, *6*(1), 40–49.

Hallgren, S. (2002). Polynomial-time quantum algorithms for pell's equation and the principal ideal problem. *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, 653–658.

Harrow, A. W., Hassidim, A., & Lloyd, S. (2009). Quantum algorithm for linear systems of equations. *Physical Review Letters*, *103*(15), 150502.

Hastie, T., Tibshirani, R., & Tibshirani, R. (2020). Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, *35*, 579–592. https://doi.org/10.1214/19-STS733

Hazimeh, H., & Mazumder, R. (2018). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, *68*. https://doi.org/10.1287/opre.2019.1919

Henriksen, L. F., & Seabrooke, L. (2016). Transnational organizing: Issue professionals in environmental sustainability networks. *Organization*, *23*(5), 722–741.

Hocking, R., & Leslie, R. (1967). Selection of the best subset in regression analysis. *Technometrics*, *9*(4), 531–540.

Høyer, P., Neerbek, J., & Shi, Y. (2002). Quantum complexities of ordered searching, sorting, and element distinctness. *Algorithmica*, *34*(4), 429–448.

Hu, J., & Wang, Y. (2020). Quantum annealing via path-integral monte carlo with data augmentation. *Journal of Computational and Graphical Statistics*, 1–13.

Huang, Y.-S., Guilleminault, C., Hwang, F.-M., Cheng, C., Lin, C.-H., Li, H.-Y., & Lee, L.-A. (2016). Inflammatory cytokines in pediatric obstructive sleep apnea. *Medicine*, *95*(41).

Humrich, J. Y., & Riemekasten, G. (2016). Restoring regulation–il-2 therapy in systemic lupus erythematosus. *Expert review of clinical immunology*, *12*(11), 1153–1160.

Jackson, S. (2020). Towards transformative solidarity: Reflections from amnesty international's global transition programme. *Emory Int'l L. Rev.*, *34*, 705.

Jansen, S., Ruskai, M.-B., & Seiler, R. (2007). Bounds for the adiabatic approximation with applications to quantum computation. *Journal of Mathematical Physics*, *48*(10), 102111.

Jing, C., Castro-Dopico, T., Richoz, N., Tuong, Z. K., Ferdinand, J. R., Lok, L. S., Loudon, K. W., Banham, G. D., Mathews, R. J., Cader, Z., et al. (2020). Macrophage metabolic reprogramming presents a therapeutic target in lupus nephritis. *Proceedings of the National Academy of Sciences*, *117*(26), 15160–15171.

Joachim, J. (2003). Framing issues and seizing opportunities: The un, ngos, and women's rights. *International Studies Quarterly*, *47*(2), 247–274.

Johns, M., & Hocking, B. (1997). Daytime sleepiness and sleep habits of australian workers. *Sleep*, *20*(10), 844–847.

Johns, M. W. (1991). A new method for measuring daytime sleepiness: The epworth sleepiness scale. *sleep*, *14*(6), 540–545.

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, *4*(1).

Jordan, L., & Van Tuijl, P. (2000). Political responsibility in transnational ngo advocacy. *World development*, *28*(12), 2051–2065.

Jordan, S. P. (2005). Fast quantum algorithm for numerical gradient estimation. *Physical Review Letters*, *95*(5), 050501.

Kahler, M. (2011). *Networked politics: Agency, power, and governance*. Cornell University Press.

Kandala, A., Temme, K., Córcoles, A. D., Mezzacapo, A., Chow, J. M., & Gambetta, J. M. (2019). Error mitigation extends the computational reach of a noisy quantum processor. *Nature*, *567*(7749), 491–495. https://doi.org/10.1038/s41586-019-1040-7

Kang, J.-H., & Lin, H.-C. (2012). Obstructive sleep apnea and the risk of autoimmune diseases: A longitudinal population-based study. *Sleep medicine*, *13*(6), 583–588.

Kassal, I., Jordan, S. P., Love, P. J., Mohseni, M., & Aspuru-Guzik, A. (2008). Polynomial-time quantum algorithm for the simulation of chemical dynamics. *Proceedings of the National Academy of Sciences*, *105*(48), 18681–18686.

Keck, M. E., & Sikkink, K. (1998). *Activists beyond borders: Advocacy networks in international politics*. Cornell University Press. http://www.jstor.org/stable/10.7591/j.ctt5hh13f

Kim, E. Y., & Moudgil, K. D. (2017). Immunomodulation of autoimmune arthritis by pro-inflammatory cytokines. *Cytokine*, *98*, 87–96.

Kishimoto, T., Kang, S., & Tanaka, T. (2015). Il-6: A new era for the treatment of autoimmune inflammatory diseases. *Innovative Medicine*, 131–147.

Kobak, D., & Linderman, G. C. (2019). Umap does not preserve global structure any better than t-sne when using the same initialization. *bioRxiv*.

Kostrzewa-Janicka, J., Śliwiński, P., Wojda, M., Rolski, D., & Mierzwińska-Nastalska, E. (2016). Mandibular advancement appliance for obstructive sleep apnea treatment. *Respiratory treatment and prevention* (pp. 63–71). Springer.

Kuwabara, T., Ishikawa, F., Kondo, M., & Kakiuchi, T. (2017). The role of il-17 and related cytokines in inflammatory autoimmune diseases. *Mediators of inflammation*, *2017*.

Kwiat, P., Mitchell, J., Schwindt, P., & White, A. (2000). Grover's search algorithm: An optical approach. *Journal of Modern Optics*, *47*(2-3), 257–266.

Lasswell, H. D. (2018). *Politics: Who gets what, when, how*. Pickle Partners Publishing.

Liu, X., Cheng, H.-M., & Zhang, Z.-Y. (2019). Evaluation of community detection methods. *IEEE Transactions on Knowledge and Data Engineering*, *32*(9), 1736–1746.

Liu, Z., Bethunaickan, R., Huang, W., Ramanujam, M., Madaio, M. P., & Davidson, A. (2011). Ifn-α confers resistance of systemic lupus erythematosus nephritis to therapy in nzb/w f1 mice. *The Journal of Immunology*, *187*(3), 1506–1513.

Lloyd, S., Mohseni, M., & Rebentrost, P. (2014). Quantum principal component analysis. *Nature Physics*, *10*(9), 631–633.

Loftus Jr, E. V. (2007). Biologic therapy in crohn's disease: Review of the evidence. *Reviews in gastroenterological disorders*, *7*, S3–12.

Long, G.-L. (2001). Grover algorithm with zero theoretical failure rate. *Physical Review A*, *64*(2), 022307.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Maoz, Z. (2017). Network science and international relations. *Oxford research encyclopedia of politics*.

Marsaglia, G., Tsang, W. W., Wang, J., et al. (2003). Evaluating kolmogorov's distribution. *Journal of Statistical Software*, *8*(18), 1–4.

Masarsky, C. S. (2018). Hypoxic stress: A risk factor for post-concussive hypopituitarism? *Medical hypotheses*, *121*, 31–34.

Mohamed Ezzat, M. H., Mohammed, A. A., Ismail, R. I., & Shaheen, K. Y. (2016). High serum april levels strongly correlate with disease severity in pediatric atopic eczema. *International journal of dermatology*, *55*(9), e494–e500.

Moloo, R. (2011). The quest for legitimacy in the united nations: A role for ngos? *UCLA Journal of International Law and Foreign Affairs*, 1–40.

Movahedi, M., Mahdaviani, S. A., Rezaei, N., Moradi, B., Dorkhosh, S., & Amirzargar, A. A. (2008). Il-10, tgf-ß, il-2, il-12, and ifn- cytokine gene polymorphisms in asthma. *Journal of Asthma*, *45*(9), 790–794. https://doi.org/10.1080/02770900802207261

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., & Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *science*, *328*(5980), 876–878.

Murdie, A. (2014). The ties that bind: A network analysis of human rights international nongovernmental organizations. *British Journal of Political Science*, *44*(1), 1–27.

Murdie, A., & Davis, D. R. (2012). Looking in the mirror: Comparing ingo networks across issue areas. *The Review of International Organizations*, *7*(2), 177–202.

Murdie, A., & Polizzi, M. (2017). Human rights and transnational advocacy networks. *The oxford handbook of political networks*.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM journal on computing*, *24*(2), 227–234.

Nelson, P. J. (1997). Conflict, legitimacy, and effectiveness: Who speaks for whom in transnational ngo networks lobbying the world bank? *Nonprofit and Voluntary Sector Quarterly*, *26*(4), 421–441. https://doi.org/10.1177/0899764097264003

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), 8577–8582.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

Nielsen, M. A., & Chuang, I. L. (2010). *Quantum computation and quantum information*. Cambridge University Press.

Ochiai, H., Shirasawa, T., Nishimura, R., Morimoto, A., Shimada, N., Ohtsu, T., Kujirai, E., Hoshino, H., Tajima, N., & Kokaze, A. (2010). Relationship of body mass index to percent body fat and waist circumference among schoolchildren in japan-the influence of gender and obesity: A population-based cross-sectional study. *BMC Public Health*, *10*(1), 493.

Oflazoglu, E., Grewal, I. S., & Gerber, H. (2009). Targeting cd30/cd30l in oncology and autoimmune and inflammatory diseases. *Therapeutic Targets of the TNF Superfamily*, 174–185.

Omachi, T. A., Claman, D. M., Blanc, P. D., & Eisner, M. D. (2009). Obstructive sleep apnea: A risk factor for work disability. *Sleep*, *32*(6), 791–798.

Oztuna, D. (n.d.). Ea (2006). investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journals of Medical Science*, *171*.

Pallas, C. L., & Urpelainen, J. (2013). Mission and interests: The strategic formation and function of north-south ngo campaigns. *Global Governance*, *19*(3), 401–423. http://www.jstor.org/stable/24526201

Phillips, B. G., Wang, Y., Ambati, S., Ma, P., & Meagher, R. B. (2020). Airways therapy of obstructive sleep apnea dramatically improves aberrant levels of soluble cytokines involved in autoimmune disease. *Clinical Immunology*, *221*, 108601.

Rajaratnam, S. M., Barger, L. K., Lockley, S. W., Shea, S. A., Wang, W., Landrigan, C. P., O'Brien, C. S., Qadri, S., Sullivan, J. P., Cade, B. E., et al. (2011). Sleep disorders, health, and safety in police officers. *Jama*, *306*(23), 2567–2578.

Rebentrost, P., Mohseni, M., & Lloyd, S. (2014). Quantum support vector machine for big data classification. *Physical Review Letters*, *113*(13), 130503.

Risse-Kappen, T., Risse, T., Ropp, S. C., Sikkink, K., et al. (1999). *The power of human rights: International norms and domestic change*. Cambridge University Press.

Rtsne, K. J. (2017). T-distributed stochastic neighbor embedding using a barnes-hut implementation. 2015. *URL https://github. com/jkrijthe/Rtsne. R package version 0.13*.

Sadiku, P., & Walmsley, S. R. (2019). Hypoxia and the regulation of myeloid cell metabolic imprinting: Consequences for the inflammatory response. *EMBO reports*, *20*(5), e47388.

Samy, E., Wax, S., Huard, B., Hess, H., & Schneider, P. (2017). Targeting baff and april in systemic lupus erythematosus and other antibody-associated diseases. *International reviews of immunology*, *36*(1), 3–19.

Schuld, M., & Petruccione, F. (2018). *Supervised learning with quantum computers*. Springer.

Schuld, M., Sinayskiy, I., & Petruccione, F. (2016). Prediction by linear regression on a quantum computer. *Physical Review A*, *94*(2), 022342.

Serebrovskaya, T. V., & Xi, L. (2015). Intermittent hypoxia in childhood: The harmful consequences versus potential benefits of therapeutic uses. *Frontiers in pediatrics*, *3*, 44.

Shen, X., Pan, W., & Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, *107*(497), 223–232.

Shokrgozar, M. A., Sarial, S., Amirzargar, A., Shokri, F., Rezaei, N., Arjang, Z., Radfar, J., Yousefi-Behzadi, M., Sahraian, M. A., & Lotfi, J. (2009). Il-2, ifn-γ, and il-12 gene polymorphisms and susceptibility to multiple sclerosis. *Journal of clinical immunology*, *29*(6), 747–751.

Shor, P. W. (1994). Algorithms for quantum computation: Discrete logarithms and factoring. *Proceedings 35th annual symposium on foundations of computer science*, 124–134.

Shor, P. W. (1999). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review*, *41*(2), 303–332.

Short, M. (2013). Improved inequalities for the poisson and binomial distribution and upper tail quantile functions. *International Scholarly Research Notices*, *2013*.

Shumate, M., & Dewitt, L. (2008). The north/south divide in ngo hyperlink networks. *Journal of Computer-Mediated Communication*, *13*(2), 405–428.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, *19*(2), 279–281.

Smirnov, N. V. (1939). Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University*, *2*(2), 3–16.

Stovel, K., & Shaw, L. (2012). Brokerage. *Annual review of sociology*, *38*, 139–158.

Stroup, S. S., & Wong, W. H. (2017). *The authority trap*. Cornell University Press.

Sullivan, B. N., & Stewart, D. (2017). Do connections always help? network brokerage's negative impact on the emergence of status. *Emergence*. Emerald Publishing Limited.

Sung, Y., Beaudoin, F., Norris, L. M., Yan, F., Kim, D. K., Qiu, J. Y., von Lüpke, U., Yoder, J. L., Orlando, T. P., Gustavsson, S., et al. (2019). Non-gaussian noise spectroscopy with a superconducting qubit sensor. *Nature Communications*, *10*(1), 1–8.

Szegedy, M. (2004). Quantum speed-up of markov chain based algorithms. *45th Annual IEEE symposium on foundations of computer science*, 32–41.

Tabarkiewicz, J., Pogoda, K., Karczmarczyk, A., Pozarowski, P., & Giannopoulos, K. (2015). The role of il-17 and th17 lymphocytes in autoimmune diseases. *Archivum immunologiae et therapiae experimentalis*, *63*(6), 435–449.

Tallberg, J., Dellmuth, L. M., Agné, H., & Duit, A. (2018). Ngo influence in international organizations: Information, access and exchange. *British journal of political science*, *48*(1), 213–238.

Tanaka, T., Narazaki, M., & Kishimoto, T. (2014). Il-6 in inflammation, immunity, and disease. *Cold Spring Harbor perspectives in biology*, *6*(10), a016295.

Terman, R. (2019). Rewarding resistance: Theorizing defiance to international shaming. *Manuscript, University of Chicago*.

Thangarajh, M., Masterman, T., Rot, U., Duvefelt, K., Brynedal, B., Karrenbauer, V. D., & Hillert, J. (2005). Increased levels of april (a proliferation-inducing ligand) mrna in multiple sclerosis. *Journal of neuroimmunology*, *167*(1-2), 210–214.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Toubi, E., & Vadasz, Z. (2019). Innate immune-responses and their role in driving autoimmunity. *Autoimmunity reviews*, *18*(3), 306–311.

Townsend, J., & Townsend, A. (2004). Accountability, motivation and practice: Ngos north and south. *Social & Cultural Geography*, *5*(2), 271–284. https://doi.org/10.1080/1464936042000169025 9

Vakil, M., Park, S., & Broder, A. (2018a). The complex associations between obstructive sleep apnea and auto-immune disorders: A review. *Medical Hypotheses*, *110*, 138–143.

Vakil, M., Park, S., & Broder, A. (2018b). The complex associations between obstructive sleep apnea and auto-immune disorders: A review. *Medical hypotheses*, *110*, 138–143.

Van Dam, W., & Shparlinski, I. E. (2008). Classical and quantum algorithms for exponential congruences. *Workshop on Quantum Computation, Communication, and Cryptography*, 1–10.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the cl common language effect size statistics of mcgraw and wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132.

Wang, G. (2017). Quantum algorithm for linear regression. *Physical review A*, *96*(1), 012335.

Wang, Y. (2011). Quantum monte carlo simulation. *The Annals of Applied Statistics*, *5*(2A), 669–683.

Wang, Y., Meagher, R. B., Ambati, S., Ma, P., & Phillips, B. G. (2020). Patients with obstructive sleep apnea have suppressed levels of soluble cytokine receptors involved in neurodegenerative disease, but normal levels with airways therapy. *Sleep and Breathing*, 1–13.

Ward, C. P., McCoy, J. G., McKenna, J. T., Connolly, N. P., McCarley, R. W., & Strecker, R. E. (2009). Spatial learning and memory deficits following exposure to 24 h of sleep fragmentation or intermittent hypoxia in a rat model of obstructive sleep apnea. *Brain research*, *1294*, 128–137.

Wasserman, S., Faust, K. et al. (1994). Social network analysis: Methods and applications.

Welch, W. J. (1982). Algorithmic complexity: Three np-hard problems in computational statistics. *Journal of Statistical Computation and Simulation*, *15*(1), 17–25.

Wiebe, N., Braun, D., & Lloyd, S. (2012). Quantum algorithm for data fitting. *Physical Review Letters*, *109*(5), 050505.

Wocjan, P., Chiang, C.-F., Nagaj, D., & Abeyesinghe, A. (2009). Quantum algorithm for approximating partition functions. *Physical Review A*, *80*(2), 022340.

Wong, W. H. (2012). *Internal affairs*. Cornell University Press.

You, T., Cheng, H.-M., Ning, Y.-Z., Shia, B.-C., & Zhang, Z.-Y. (2016). Community detection in complex networks using density-based clustering algorithm and manifold learning. *Physica A: Statistical Mechanics and its Applications*, *464*, 221–230.

Yuan, M., & Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 143–161.

Yuan, X.-T., Li, P., & Zhang, T. (2017). Gradient hard thresholding pursuit. *The Journal of Machine Learning Research*, *18*(1), 6027–6069.

Zar, J. H. (1972). Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, *67*(339), 578–580.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942.

Zhang, J., & Cao, J. (2017). Finding common modules in a time-varying network with application to the drosophila melanogaster gene regulation network. *Journal of the American Statistical Association*, *112*(519), 994–1008.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, *7*(Nov), 2541–2563.

Zhao, Z., Fitzsimons, J. K., & Fitzsimons, J. F. (2019). Quantum-assisted gaussian process regression. *Physical Review A*, *99*(5), 052331.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, *67*(2), 301–320.