Research Statement

Shawn Mankad
Cornell University

1. Introduction

The rise of analytics and big data has impacted our daily lives and shaped the production and provision of goods and services. Data analytics also poses new opportunities and challenges for researchers. It is easier than ever to develop and test theory by collecting data over time or in different conditions at the most granular levels. On the other hand, the structure of modern data creates modeling and statistical challenges that must be addressed for the promise of big data and analytics to be achieved. For example, text data is ubiquitous in the form of online reviews, corporate filings, service call transcripts, and so on. Yet, simply summarizing large volumes of text data, let alone performing causal inference with it, is a difficult problem and active area of research in several domains. Moreover, these methodological issues have system-design implications. For instance, the ability to use text analysis methods to identify the author of anonymous writing has enormous privacy implications. Similarly, data is readily available on the individual components of many complex systems, creating network data (e.g., social networks, financial networks, retail-store networks), which feature unique statistical issues (e.g., reverse causality, community structure) that stand in the way of gaining insight into many real-world business and policy problems.

The goal of my research agenda is to address these issues and help solve business, economic, and policy problems through the development and application of the theory and methods of statistics and data science. My research has two major themes. From a statistical point of view, one theme is on methodology for network and text data, which include matrix factorization and mixture models. From an application areas point of view, I have gained extensive experience and expertise in domain problems like measuring service quality in text data, extracting information in online reviews, and measuring the interconnectedness of financial firms.

My research also has had impact beyond academia, having been featured in the Harvard Business Review and numerous media outlets such as the Wall Street Journal, Bloomberg, Chicago Tribune, and so on. My work on developing methods for assessing the interconnectedness of financial institutions and systemic risk has been utilized in operations at the Federal Reserve Board and is also supported by funding from the National Science Foundation.

Below I summarize my methodological research published in the Statistics domain, followed by a summary of applied work published in the management sciences.

2. Research in Applied Statistics and Data Science: Multivariate Analysis Methods

From a methodology perspective, a recurring theme in several of my papers is that of matrix factorization and multivariate methods. A key idea in my work is often to extend and customize a matrix factorization model by incorporating contextual information into the constraints that define the factorization. This leads to optimization or estimation challenges that I have overcome using a variety of techniques, including Bayesian modeling [21], gradient-descent techniques [2, 4, 10], and other heuristics [1, 3, 9].

Next, I summarize several papers that develop variants of the non-negative matrix factorization for modeling of different data, from dynamic networks (where the network observations are evolving over time) and text documents over time, to stock returns.

In [2], we describe the development and application of a smoothed non-negative matrix factorization for exploration and time-varying community detection in time-evolving graph sequences. The matrix factorization model allows the user to home in on and display interesting, underlying structure and its evolution over time. The methods are scalable to weighted networks with a large number of time points or nodes, and can accommodate sudden changes to graph topology. Our techniques are demonstrated with several dynamic graph series from both synthetic and real-world data, including citation and trade networks. These examples illustrate how users can steer the techniques and combine them with existing methods to discover and display meaningful patterns in sizable graphs over many time points.

In [4], we extend the paper above to create a smoothed or semi-supervised factorization that emphasizes analyst specified network motifs when searching for community structure and influential nodes. The ability for the analyst to emphasize particular nodes with precalculated centrality scores allows the model to be customized to different contexts, and therefore bridges the numerous and separate approaches that have been developed for different settings (e.g., social versus banking networks). The model is used to investigate whether Twitter conversations between legislators reveal their real-world position and influence by analyzing multiple Twitter networks that feature different types of link relations between the Members of Parliament (MPs) in the United Kingdom. Leveraging only link relation data, we find that important politicians in Twitter networks are associated with real-world leadership positions, and that rankings from the proposed method are correlated with the number of future media headlines.

In [10], we develop a new variant of semi-non-negative matrix factorization for software developers wanting to use reviews of their mobile app to inform future development efforts. This paper uses real-world considerations to motivate the model: constrained factorization is used to perform topic and sentiment modeling of text, and this factorization is embedded within an ordered logistic regression to account for the structure of online reviews. By performing the entire estimation in a single optimization framework, we show that the model produces more accurate out-of-sample rating predictions based on the text. Projected gradient descent methods are adapted for estimation. From a business point of view, this paper shows how service quality information within the text can be identified and operationalized. We apply our approach to systematically compare different apps over time for benchmarking of features and consumer sentiment. The dataset consists of over 100,000 mobile reviews over several years for three of the most popular online travel agent apps from the iTunes and Google Play marketplaces.

In [21], we propose a new approach to estimate the portfolio composition of banks as function of their daily trading activity and stock returns. From the estimated portfolio composition, we then derive an index of portfolio concentration (bank-specific risk) for each individual bank and an index of portfolios similarity across banks (systemic risk) which captures market susceptibility to propagating shocks to any asset class. The paper has several contributions. From a methods point of view, the proposed technique involves solving a matrix factorization problem within a novel Bayesian estimation framework for the matrix factorization area. It has substantially better estimation properties than estimates obtained from more commonly used numerical optimization techniques. From a managerial point of view, the paper offers a new set of systemic risk indicators that arguably capture the potential of contagion due to common asset holdings more cleanly than competing measures. In fact, we find evidence that systemic risk measures derived from our approach lead, in a forecasting sense, commonly used systemic risk indicators.

In [20], we develop and present a novel methodology to detect regime changes within a sequence of networks that have overlapping and evolving community structure. The core of the methodology is a non-negative matrix factorization that maximizes a Poisson likelihood subject to a penalty that accounts for sparsity in the network. By fitting the factorization model over a rolling window with a fast-numerical optimization algorithm, change detection is accomplished by statistical monitoring of the matrix factors' evolution. A novel statistic is used to characterize the overall network evolution as well as the contribution of each node to the change. We demonstrate that the proposed methodology compares favorably with alternative techniques for on-the-go network change detection using synthetic and real data. A detailed case study on the 2007-2009 financial crisis and the European sovereign debt crisis shows the promise of the methodology for regulators as it identifies particular banks that contributed to each crisis in addition to identifying changing market conditions. The sole co-author on this paper is a 2nd year PhD student.

Moving beyond matrix factorization, I have also explored statistical modeling of dynamic network using stochastic processes. I highlight two of my papers in this stream.

In [15], we consider the problem of change point detection for networks, which can support formal monitoring and early-warning systems to aid regulators and market participants. We propose new methodology for change detection in attributed and sparse banking networks that detects changes substantially faster than alternative change detection approaches. They key innovation is to utilize Hurdle models to account for sparsity (lack of edge existence) when modeling the network structure. The proposed methodology combines state space models, Hurdle models, and control charts, and would have raised alarms to regulators prior to several key events leading to the 2007-2009 financial crisis.

A key question for social media platforms is to determine influential users, in the sense that they generate interactions between members of the platform. Common measures used both in the academic literature and by companies that provide analytics services are variants of the popular web search PageRank algorithm applied to networks that capture connections between users. In [7], we develop an alternative modeling framework using multivariate interacting counting processes to additionally capture the detailed actions that users undertake on such platforms, namely posting original content, reposting and/or mentioning other users' postings. Based on the proposed model, we also derive a novel influence measure. We discuss estimation of the model parameters through maximum likelihood and establish their asymptotic properties. The proposed model and the accompanying influence measure are illustrated on a dataset covering a five-year period of the Twitter actions of the members of the U.S. Senate, as well as mainstream news organizations and media personalities.

Motivated by the Operations Management and franchising literature, I have also extended econometric techniques to estimate the network itself, where edges are defined as the causal impact one store's performance has on another store. In [22], we consider a network of stores operating under the same brand, for example, a chain of coffee shops or banks. Increasing sales at one store may have a different impact on the sales of another store: from a negative impact as a result of potential cannibalization to a positive impact from increased brand awareness and customer engagement. We study how to causally identify the spatially heterogeneous network effects between store pairs, which can be used to measure store influence. The novel and unique feature of our approach is an extension of a spatial econometrics model that allows for identification of both positive and negative peer effects between stores. This semi-parametric methodology allows us to handle a large-scale network and provides causal estimates for the network effects between all store pairs. Our influence estimates provide valuable insights for the design and management of networks; for example, for prioritizing stores with the highest influences for ownership or improvement to optimize return on investment.

I also have published research relating to classical Statistics, such as inverse problems and isotonic (monotonic) regression. In [5, 6], we investigate two-stage plans based on nonparametric procedures for estimating an inverse regression function at a given point. Specifically, isotonic regression is used at stage one to obtain an initial estimate followed by another round of isotonic regression in the vicinity of this estimate at stage two. It is shown that such two-stage plans accelerate the convergence rate of one-stage procedures and are superior to existing two-stage procedures that use local parametric approximations at stage two when the available budget is moderate and/or the regression function is "ill-behaved."

3. Research in IS/OM: Text Analytics for Online Reviews and Digital Platforms

Online review platforms like Yelp and Amazon have become a major topic of study at the interface of Information Systems, Operations Management, and Marketing. Accordingly, there is an extensive literature showing that online reviews are useful for understanding customer engagement, service and product quality, and firm-level performance measures. The two most common elements used in prior works are the star rating of the online review, typically on a 1-5 scale, and the number of posted reviews, representing traffic or popularity of the provider – note that the textual content within an online review has been underutilized historically. Next, I summarize several of my papers that focus on the review text as a valuable source of additional information to sharpen previous findings, explore new research questions, and expand the potential use-cases of such platforms.

Measuring service quality remains a challenging problem given that traditional methodologies (based on surveys, field data, etc.) are often costly and hard to perform at scale across the competitive landscape. Further, understanding how elements of service quality contribute to the performance of service providers continues to be a relevant problem for the service industry. My papers address these challenges in the context of the hotel and restaurant sector, vital components of the service economy.

My earliest work on this problem in [8] applies the Latent Dirichlet Allocation model and related natural language processing methods to illustrate how review text can be automatically evaluated. This paper was among the first in the Service Operations Management literature to utilize methods from computational linguistics. Our analysis shows that longer, more focused reviews tend to have a higher negative sentiment. Reviews that talk about experience tend to be positive while reviews that talk about transactions or value tend to be negative. Using these observations, managers of hotel operations can develop an effective strategy for responding to customer feedback and significantly improving their service offerings.

Adding rigor to the validation and interpretation of the results in addition to advancing the above methodology, in [14], we investigate Aspect Based Sentiment Analysis methods to simultaneously estimate topics and the sentiment along each dimension within each review. We find that our extracted service quality dimensions match industry gold standards and are correctly identified by subjects in a controlled laboratory setting. Furthermore, we show that specific service dimensions are significantly correlated with the survival of merchants, even after controlling for competition and other factors. This has important implications: our work provides a scalable methodology for using the text in social media to measure the quality of service providers associated with economic outcomes for the providers. Our method requires much lower human effort compared to classical survey-based approaches.

I have also examined broader policy issues involving digital platforms pertaining to moral hazard in hygiene inspections and user-privacy. In [12], we study how online reviews can be used to tackle the policy-oriented and socially relevant problem of moral hazard within the New York City restaurant hygiene inspection apparatus. While health inspection programs are designed to protect consumers, such inspections typically occur at wide intervals of time, allowing restaurant hygiene to remain unmonitored in the interim periods. To the extent that social media provides some visibility into the hygiene practices of restaurants, we argue that the effects of information asymmetry that lead to moral hazard may be partially mitigated in this context. Using a dataset of restaurant hygiene inspections in New York City from 2010 through 2016, and the associated set of online reviews for the same set of restaurants from Yelp, this work combines supervised learning techniques with natural language processing methods to build a custom dictionary that can be used to measure the hygiene quality of restaurants in real-time with online reviews. Our contribution to the literature includes the dictionary as a research artifact, empirical insights created as a result of combining machine learning and econometric methods, and providing a clear example of how online reviews can be used to partially overcome moral hazard, a classical issue in economics.

In [17], we are among the first to study the lack of anonymity in online reviews – a key issue at the very core of online review platforms. Although incentivizing high-quality reviews is an important business objective for the platform, we show that it is also possible to identify anonymous posters by exploiting the characteristics of posted content. We present a novel two-stage authorship attribution methodology that combines structured and text data by identifying an author first by the amount and granularity of structured data (e.g., location, first name) posted with the online review and second by the author's writing style. As a case study, we show that 75% of the 1.3 million users in data publicly released by Yelp are uniquely identified by three structured variable combinations. For the remaining 25%, when the number of potential authors with (nearly) identical structured data ranges from 100 to 5 and sufficient training data exists for text analysis, the average probabilities of identification range from 40 to 81%. Our findings suggest that platforms concerned with the potential negative effects of privacy-related incidents should limit or generalize their posters' structured data when it is adjoined with textual content or mentioned in the text itself. We also show that although protection policies that focus on structured data remove the most predictive elements of authorship, they also have a small negative effect on the usefulness of content.

In [23], we develop a Bayesian data protection model to synthetically alter users' structured data based on textual content in their posted content with a single hyperparameter designed for privacy. We demonstrate the method using review data from an online platform to evaluate the effects of protection policies that the platform or the user can pursue to decrease the chances of de-anonymization when anonymity is desired. We show that platform protection policies that alter the structured data with the Bayesian model removes the most predictive elements of authorship, while retaining the essential content.

4. Research in Finance: Network Analytics to Measure the Interconnectedness of Financial Institutions

Supported by a National Science Foundation grant and collaborations with scholars at the Federal Reserve Board, my research is aimed at developing new techniques that further our understanding of the interconnectivity of financial institutions. Specifically, my work focuses on the development and application of network modeling techniques that create different views of the financial system. These views ultimately facilitate a deeper understanding of underlying market dynamics, relationships between firms, and thus a more accurate assessment of vulnerabilities of the broader financial system. Below I highlight papers of mine in this stream.

Broadly speaking, the literature considers two types of financial networks: (i) "physical networks", which are composed of transactional relationships (e.g., borrowing/lending or future payout obligations via a credit default swap), and (ii) "correlation networks", which are composed of statistical relationships, inferred from stock returns, and thus driven fundamentally by overlapping portfolios of common asset holdings. In [13], we unify both types of networks together using an accounting framework for the banking sector. The paper also empirically studies both types of networks around the 2008 financial crisis. While the two networks behave similarly pre-crisis, during the crisis the correlation network shows an increase in interconnectedness, while the physical network highlights a marked decrease in interconnectedness. Importantly from a policy perspective, we find that these networks respond differently to monetary and macroeconomic shocks. Physical networks forecast liquidity problems, while correlation networks forecast financial crises.

In [18], we study the motivations of traders in the interbank market around the 2007-09 subprime crisis. We extend a commonly used statistic called market Sidedness to a panel setting to study the dispersion of beliefs for banks domiciled in different European countries. We find that country-level Sidedness reveals information from the interbank market: Sidedness leads sovereign CDS spreads and reacts to central bank interventions introduced during the crisis. Our results map the linkages between the interbank market and sovereigns as well as shed light on the channels that give rise to the sovereign-bank nexus (the set of connections linking sovereigns and the banking sector as a conduit for transmitting distress between the two).

In [19], we propose a new directed network construct—the liquidity network—to capture the urgency to trade by connecting the initiating party in a trade to the passive party. Alongside the conventional trading network connecting sellers to buyers, we show that the aggressiveness to trade reflected in the liquidity network improves short-term forecasts of soft information and country-specific yield spreads, settings where asymmetric information is likely to be more pronounced. Our work contributes to a better understanding of how interbank markets operate and convey information about the real economy. While other papers focus on interbank network structures and contagion, we focus on different network constructs and whether viewing interbank networks under different lenses might provide important insights into the macroeconomy.

5. Policy Research with Econometrics and Causal Inference Techniques

I have also developed expertise in econometrics and causal analysis, which is critical for informing policy and managerial decision making. Next, I summarize my works that aim to influence policy.

Minimum wage is an important yet controversial topic that has received attention for decades. Our study is the first to take an operational lens and empirically study the impact of the minimum wage on firms' scheduling practices. In [24], we empirically identify and highlight a new operational mechanism through which increasing the minimum wage may negatively impact worker welfare. Our analysis suggests that the combination of the reduced hours, lower eligibility for benefits, and less consistent schedules (that resulted from the minimum wage increase) may substantially hurt worker welfare, even when the overall employment at the stores stay unchanged. By better understanding the intrinsic trade-off of firms' scheduling decisions, policy makers can better design minimum wage policies that will truly benefit workers. My contribution to this research collaboration is in the empirical design, methods, and implementation. The paper utilizes difference-in-difference within a non-conventional setting with multiple treatment levels at different times across the experimental group. I also executed a majority of the data management, statistical summaries, and model estimation. A version of this paper for the general public was published in the Harvard Business Review [16] and covered extensively in the public press (Wall Street Journal, Bloomberg, etc.).

In [11], we investigate how financial regulations are formed. Specifically, we examine the text documents generated through the rule-making process to gain insight into how new regulatory regimes are implemented following major laws like the landmark Dodd-Frank Wall Street Reform and Consumer Protection Act. Due to the variety of constituent preferences and political pressures, we find evidence that the government implements rules strategically to extend the regulatory boundary by first pursuing procedural rules that establish how economic activities will be regulated, followed by specifying who is subject to the procedural requirements. Our findings together with the unique nature of the Dodd-Frank Act translate to a number of stylized facts that should guide development of formal models of the rule-making process.

6. Conclusion

My early research focused on developing multivariate methods for pattern extraction in different types of three-way arrays. The objective of that research was to improve clustering and data exploration of networks or similarly structured data. In recent research, I have transitioned to focus on solving real-world business and policy problems, and my work has featured a blend of econometrics, modeling, and advanced statistical methods for text and network analysis. The publication venues for my work have also shifted accordingly. My early research was published exclusively in methodology and Statistics journals, whereas my more recent work has been published in journals within the management sciences. Going forward, I plan to continue to pursue an interdisciplinary research agenda that is methodologically focused and systematically driven by large business, economic, and policy problems that when (partially) addressed can lead to better outcomes for society.

7. References

[1] Mankad, S., Michailidis, G., & Kirilenko, A. (2013). Discovering the ecosystem of an electronic financial market with a dynamic machine-learning method. Algorithmic Finance, 2(2), 151-165.

[2] Mankad, S., & Michailidis, G. (2013). Structural and functional discovery in dynamic networks with non-negative matrix factorization. Physical Review E, 88(4), 042812.

[3] Mankad, S., & Michailidis, G. (2014). Biclustering three-dimensional data arrays with plaid models. Journal of Computational and Graphical Statistics, 23(4), 943-965.

[4] Mankad, S., & Michailidis, G. (2015). Analysis of multiview legislative networks with structured matrix factorization: Does Twitter influence translate to the real world? The Annals of Applied Statistics, 9(4), 1950-1972.

[5] Mankad, S., Michailidis, G., & Banerjee, M. (2015). Threshold Value Estimation Using Adaptive Two-Stage Plans in R. Journal of Statistical Software, 67(1), 1-19.

[6] Tang, R., Banerjee, M., Michailidis, G., & Mankad, S. (2015). Two-stage plans for estimating the inverse of a monotone function. Technometrics, 57(3), 395-407.

[7] Xia, D., Mankad, S., & Michailidis, G. (2016) Measuring Influence of Users in Twitter Ecosystems Using a Counting Process Modeling Framework. Technometrics, 58:3, 360-370, DOI: 10.1080/00401706.2016.1142906

[8] Mankad, S., Han, H. S., Goh, J., & Gavirneni, S. (2016). Understanding online hotel reviews through automated text analysis. Service Science, 8(2), 124-138.

[9] Marino, S., Gideon, H. P., Gong, C., Mankad, S., McCrone, J. T., Lin, P. L., & Kirschner, D. E. (2016). Computational and empirical studies predict Mycobacterium tuberculosis specific T cells as a biomarker for infection outcome. PLoS computational biology, 12(4), e1004804.

[10] Mankad, S., Hu, S., & Gopal, A. (2018). Single stage prediction with embedded topic modeling of online reviews for mobile app management. The Annals of Applied Statistics, 12(4), 2279-2311.

[11] Mankad, S., Michailidis, G., & Kirilenko, A. (2018). On the Formation of Dodd-Frank Act Derivatives Regulations. PLoS One, 14(3), p.e0213730.

[12] Mejia, J., Mankad, S., & Gopal, A. (2019). A for Effort? Using the Crowd to Identify Moral Hazard in New York City Restaurant Hygiene Inspections. Information Systems Research, 30(4), 1363-1386.

[13] Brunetti, C., Harris, J.H., Mankad, S., & Michailidis, G. (2019). Interconnectedness in the interbank market. Journal of Financial Economics, 133(2), 520-538

[14] Mejia, J., Mankad, S., & Gopal, A. (2020). Service Quality Using Text Mining: Measurement and Consequences. Manufacturing & Service Operations Management

[15] Ebrahimi, S., Reisi-Gahrooei, M., Paynabar, K. and Mankad, S. (2020). Monitoring sparse and attributed networks with online Hurdle models. IISE Transactions, pp.1-14.

[16] Yu, Q., Mankad, S., & Shunko, M. (2021). "When a Higher Minimum Wage Leads to Lower Compensation", Harvard Business Review (digital article)

[17] Schneider, M., Mankad, S. (2021). A Two-Stage Authorship Attribution Method using Text and Structured Data for De-Anonymizing User Generated Content. Customer Needs and Solutions, 1-18. DOI: 10.1007/s40547-021-00116-x

[18] Brunetti, C., and Harris, J.H. & Mankad, S. (forthcoming). Sidedness and the Urgency to Borrow in the Interbank Market. Journal of Financial Markets.

[19] Brunetti, C., and Harris, J.H. & Mankad, S. (2021). Liquidity Networks, Interconnectedness, and Interbank Information Asymmetry.

[20] Ma, D., & Mankad, S. (2021). CP-Squared: A Method for Change-Point Detection in Core-Periphery Networks.

[21] Mankad, S., Brunetti, C., and Harris, J.H. & (2021). Bayesian Semi-Non-Negative Matrix Factorization: A Technique for Estimating Bank Holdings and Systemic Risk.

[22] Mankad, S., Shunko, M., & Yu, Q. (2021). Too Close for Comfort? Understanding Peer Effects in Large Franchised Networks.

[23] Schneider, M., Hu, M., & Mankad, S. (2021). Protecting the Anonymity of Online Users Using Prior Distributions on their Textual Content.

[24] Yu, Q., Mankad, S., & Shunko, M. (2021b). Evidence of the Unintended Scheduling Implications of Minimum Wage.