Collective Behavior over Social Networks with Data-driven and Machine Learning Models

by

Yan Leng

B.S., Beijing Jiaotong University (2013)
M.S., Massachusetts Institute of Technology (2016)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author .....
Program in Media Arts and Sciences,
School of Architecture and Planning
May 5, 2020

Certified by .....
Alex Pentland
Professor of Media Arts and Science, MIT
Toshiba Professor
Thesis Supervisor

Accepted by .....
Tod Machover
Academic Head, Program in Media Arts and Sciences

Collective Behavior over Social Networks with Data-driven and Machine Learning Models

by

Yan Leng

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on May 5, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences

Abstract

Individuals form network connections based on homophily; individuals' networks also shape their actions. Pervasive behavioral data provides opportunities for a richer view of the decisions on networks. Yet, the increasing volume, complex structures, and dynamics of behavioral data stretch the limit of conventional methods. I develop mathematical modeling (e.g., machine learning, game theory, and network science) and large-scale behavioral data to study collective behaviors over social networks. My dissertation will tackle this area in four directions, revolving around the intricate linkage between individuals' characteristics, actions, and their networks. First, I empirically investigate how social influence spreads over networks using two massive cell phone data, and theoretically model how do individuals aggregate information from local neighbors. Second, I study how to leverage influential nodes for selective network interventions (e.g., marketing and political campaigns), by proposing a centrality measure going beyond network structures. Third, I build a geometric deep learning model to infer individual preferences and make personalized recommendations to utilize noisy network information and nodal features effectively. Last, given that the network is essential, I develop a framework to infer the network connections based on observed actions, when networks are unavailable. My thesis provides building blocks for further network-based machine learning problems integrating nodal heterogeneity and network structures. Moreover, the findings on human behaviors and frameworks developed in my thesis shed light on marketing campaigns and population management.

Thesis Supervisor: Alex Pentland
Title: Professor of Media Arts and Science, MIT
Toshiba Professor

Acknowledgments

I am extremely fortunate to have interacted with and learned from so many amazing individuals, without whom my academic journey would not have gone this far. This acknowledgment only touches the surface.

First and foremost, I am indebted to my advisor Alex 'Sandy' Pentland, who has been extremely helpful for my professional development. His vision, wisdom, creativity, enthusiasm, and professionalism have been constant sources of inspiration ever since the day I took the courage to email him. Sandy is an incredibly supportive and down-to-earth advisor. Whenever I bounced ideas with him, he can always connect the dots and shape my immature idea in a more prom ming direction.

I am also grateful to my committee member, Dean Eckles. Every time I interact with Dean, I learn something new. Dean is extremely well-read and knowledgeable. His infectiously cheerful personality and his generosity with his time have been hugely helpful for me. I look up to him for his passion for social network and dedication to causal inference.

I am grateful for my committee member, Esteban Moro. Esteban has been an excellent collaborator and committee member: his insights have had a significant influence over my research, and it was his generous encouragement that got me finished my first project of in Ph.D. Esteban has also been a great mentor. His professionalism has hugely influenced and will continue to affect my academic career.

I am also grateful and fortunate to have Michael Bronstein in my committee. It was his seminal work on deep geometric learning that introduced me to this field. I am grateful for Michael's generosity of time and inspiring discussions on ideas. I look forward to continuing to learn from him and collaborate more in the future.

I am also indebted to my former advisors, Haris Koutsopolous, Jinhua Zhao, Larry Rudolph, and Xuedong Yan, who took me as a research assistant, introduced me to transportation and mobility research and provided continual support before I start Ph.D.

MIT, and especially the Media Lab, has broadened my eyesight and provided me

with the ideal environment for interdisciplinary research. I have learned from many other faculty members and students here. I have been so fortunate to have interacted with faculties from different departments to learn from their expertise. I would like to especially thank Kent Larson for allowing me to join the Andorra project, where I meet many passionate researchers with different expertise.

I wouldn't have survived graduate school without my wonderful peers. There are so many people I should extend my gratitude. My close collaborator, Xiaowen Dong, taught me a lot about research and is always there as a close friend. My office mates, Abdullah Almaatouq and Alejandro Noriega, always impressed me with their positivity and passion for making a societal impact. I am lucky to have Morgan Frank and Mohsen Mosleh, who I chatted a lot to go through the stressful time on the job market. I am also fortunate to have Michiel Bakker as a good friend, whose sense of humor has influenced me a lot. I would also like to thank my lab mates Yuan Yuan, Mohsen Bahrami, Zivvy Epstein, Eaman Jahani, Dhaval Adjodah, Tara Sowriraja, Dan Calacci, Martin Saveski, for our discussion of ideas and for friendships that made my overall graduate school experience.

Finally, I want to express my deep and sincere gratitude to my family for their continuous and unparalleled love, help, and support. As the only child, my parents always try to give me the best they can offer. It is their encouragement and sacrifice that take me this far. Last but not least, I want to thank my husband, Siyuan Liu. Siyuan is not only my partner but also a role model that I always look up to. Meeting you is the best part of my graduate school, and you excel in every part of our life.

This doctoral thesis has been examined by the following committee members:

Professor Alex 'Sandy' Pentland .....
Thesis Supervisor
Professor of Media Arts and Science, MIT
Toshiba Professor
Media Lab Entrepreneurship Program Director

Professor Michael Bronstein .....
Thesis Committee
Professor at the Department of Computing, Imperial College London

Professor Dean Eckles .....
Thesis Committee
KDD Career Development Professor in Communications and Technology
Associate Professor of Marketing at MIT Sloan School of Management

Professor Esteban Moro .....
Thesis Committee
Visiting Professor, MIT Media Lab
Associate Professor at the Universidad Carlos III de Madrid

Contents

1 Introduction **25**

1.1 Background 26

1.2 Research questions 27

2 Long range social influence in phone communication network **31**

2.1 Introduction 32

2.2 Literature review 35

2.2.1 Contagion models 35

2.2.2 Observational learning and word-of-mouth effect 36

2.3 Behavioral matching framework 36

2.3.1 Setting 37

2.3.2 Matching framework 40

2.4 Long range of social influence 42

2.5 Bayesian learning model and results 48

2.5.1 Bayesian learning model 48

2.5.2 Comparisons between the Bayesian learning model with existing models 52

2.5.3 Prediction results 53

2.6 Managerial implications 54

2.7 Discussion 55

3 Contextual centrality: going beyond network structure **61**

3.1 Introduction 62

3.2 Contextual centrality 64
3.3 Results 69
3.3.1 Methods 69
3.4 Discussion 79
3.5 Properties of contextual centrality 81
3.5.1 Bounds and distribution of contextual centrality in terms of spreadability 81
3.5.2 Robustness of contextual centrality in response to perturbations in $\mathbf{y}$ 82
3.5.3 Theoretical results of contextual centrality for Erdos-Renyi networks 84
3.5.4 The relationship between contextual centrality and other centrality measures 90
3.5.5 Relationship between approximated cascade payoff and cascade payoff 92
3.5.6 Game-theoretic interpretation of contextual centrality with local interactions 92
3.5.7 Differences between contextual centrality and centrality measures developed on weighted networks 94
3.6 Additional results for empirical analysis 96
3.6.1 Predictive power of contextual centrality in eventual adoptions 96
3.6.2 Performance relative to other centrality measures on random networks 98
3.6.3 Average approximated cascade payoff for contextual centrality and the variations of other centrality measures 102
3.6.4 Comparison of seeding strategies when $\overline{\mathbf{y}}(\mathbf{U}_1^T\mathbf{y}) < 0$ 104

4 Recommender systems with heterogeneous information: A geometric deep learning approach 107

4.1 Introduction 108

4.2 Literature review . . . 113
4.2.1 Recommender systems . . . 113
4.2.2 Geometric deep learning . . . 114
4.3 Data description . . . 116
4.3.1 Business . . . 116
4.3.2 Users . . . 118
4.4 Model . . . 120
4.4.1 Problem formulation . . . 121
4.4.2 Local smoothness regularization with graph attention network 122
4.4.3 Framework . . . 125
4.5 Results . . . 127
4.5.1 Experimental setting . . . 127
4.5.2 Learning performance . . . 129
4.5.3 Pattern analysis on business representations . . . 132
4.6 Conclusion and managerial implications . . . 135

5 Learning Quadratic Games on Networks 139

5.1 Introduction . . . 140
5.2 Network games of linear-quadratic payoffs . . . 142
5.3 Learning games with independent marginal benefits . . . 145
5.3.1 Learning framework . . . 145
5.3.2 Learning algorithm . . . 147
5.4 Learning games with homophilous marginal benefits . . . 147
5.4.1 Learning framework . . . 148
5.4.2 Learning algorithm . . . 149
5.5 Experiments on synthetic data . . . 149
5.5.1 Comparison of learning performance . . . 151
5.5.2 Learning performance with respect to different factors in network games . . . 152
5.5.3 Learning the marginal benefits . . . 154

5.6 Experiments on real world data 154
5.6.1 Social network 155
5.6.2 Trade network 155
5.7 Discussion 157

6 Conclusion **165**
6.1 Summary 165
6.2 Future work 169

A Interpretable Stochastic Block Influence Model: measuring social influence among homophilous communities **187**
A.1 Introduction 188
A.2 Related literature 191
A.3 Methodology 192
A.3.1 Stochastic Block Influence Model 192
A.3.2 Generative process 195
A.4 Experiments 198
A.5 Applications and future works 206

List of Figures

2-1 **An illustration of initial adopters, information cascade, and the hop indexes.** In the information cascade $\mathcal{C}_T$ shown in the figure, within an observation period T, the initial adopter Anne (colored green) passes information to her neighbors, Bob, Eva, Cathy, and Daniel, who after receiving information from Anne continue to pass the information onwards. Labeled with hop index one, they further diffuse the information to Franklin, Greg, Helen, Isabel, Jack, and Kate, who are then on hop index two. The process continues until the end of the observation period. Among the people who receive information, Bob, Isabel, and Daniel (colored blue) decided to adopt the behavior, while others (colored grey) decided not to. 40

2-2 **Two types of mobility frequency patterns during weekends, revealing different individual preferences: (a) an explorative pattern; (b) an exploitative pattern.** The intensity of the color represents the normalized visitation frequency, i.e., the darker red color corresponds to a more frequent visit. 41

2-3 **Percentage improvement in adoption rate relative to the control group due to social influence via phone communication network (ΔA_h)**: (a) attending cultural performance; (b) visiting the retail store. The y-axis is the difference in adoption likelihood of the two groups, and the x-axis is the hop index. The purple, blue, and red dashed lines show the estimated effect of social influence using PSM, random matching, and PSM after a shuffling test, respectively. The shaded regions correspond to the 5% and 95% confidence intervals from bootstrap sampling. The higher and lower end of the vertical line indicates the 5% and 95% interval. 44

2-4 **Matching on behavioral covariates, and on both behavioral covariates and socio-demographics, for the adoption behavior of visiting the retail store.** The y-axis is the difference in adoption likelihood of the two groups, and the x-axis is the hop index. The bar plot and the vertical lines correspond to the mean, 5%, and 95% confidence intervals, respectively. The blue and red bars correspond to behavioral matching, and behavioral + socio-demographics matching. 45

2-5 **SMD for the matching between the control group and the different treatment groups (different hop indexes h), in the case of attending cultural performance.** 46

2-6 **SMD for the matching between the control group and the different treatment groups (different hop indexes h), in the case of visiting retail store.** 47

2-7 **Percentage improvement in adoption rate relative to the control group due to social influence estimated by post-Lasso logistic regression for different treatment groups (ΔA_h)**, in the case of a) attending cultural performance and b) visiting the retail store. The vertical bars cover 5% and 95% confidence intervals. 48

2-8 **Decision-making process for Greg, according to the proposed Bayesian learning model.** At the time instance $t = 0$, Greg forms a prior understanding of the product. At $t = 1$, Bob told Greg about his evaluation of the product. Knowing Bob's general preference, Greg then updated his perception ($P_1(\boldsymbol{w}_{\text{Greg}})$). The same updating process happens after observing the preferences and the evaluations of Franklin and Helen afterward. With this illustration, we show how Greg updates his perception about the product by dynamically aggregating local information from his neighbors who communicated with him. 51

2-9 **Higher-order social influence.** Existing contagion models assume that individual behaviors are independent of the decisions of others conditioned on their immediate neighbors. That is to say, existing contagion models do not distinguish scenarios A and B in terms of the decision-making of Greg. Our model, thanks to the higher-order social influence and propagation of information between neighbors, can separate the two scenarios. 53

2-10 **Performance of different models in predicting adoption behavior: (a) attending cultural performance; (b) visiting retail store. The error bars correspond to the 5% and 95% confidence intervals.** 54

2-11 **Percentage of individuals (a,c) as well as their adoption rates (b,d) at each hop, computed using information cascades from all observation periods.** 58

3-1 **Predictive power of contextual centrality. We show how the average centrality of first-informed individuals predicts the eventual adoption rate of non-first-informed individuals in (a) microfinance and (b) weather insurance.** The y-axis shows the 95% confidence interval of R^2 computed from 1000 bootstrap samples from ordinary least squares regressions controlling for village size. The x-axis shows varying values for $p\lambda_1$, which influences only diffusion centrality and contextual centrality. 72

3-2 **Performance of contextual centrality relative to other centrality measures on random networks.** Each plot shows the relative change, computed as $\frac{a-b}{\max(|a|,|b|)}$ where a is CC's average payoff and b is the maximum average payoff of the other centrality measures, for varying values of $\frac{\overline{\mathbf{y}}}{\sigma(\mathbf{y})}$ and $p\lambda_1$. The plots correspond to the results on random networks generated according to the (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 74

3-3 **Average payoffs when standardized average contribution is 0.** Here we show the average payoff with 95% confidence interval when seeding with different methods on (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 75

3-4 **Average payoffs when standardized average contribution is 1.** Here we show the average payoff with 95% confidence interval when seeding with different methods on (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 76

3-5 **Performance of contextual centrality relative to other centrality measures on real-world networks, including (a) microfinance, (b) weather insurance, and (c) political campaign.** Each plot shows the relative change for varying values of $p\lambda_1$. We compare contextual centrality with degree centrality, diffusion centrality, eigenvector centrality, Katz centrality, and random seeding. 77

3-6 **Average cascade payoff for variations of contextual centrality and eigenvector centrality. The x-axis is $p\lambda_1$, and the y-axis is the average payoff, with the shaded region as the 95% confidence interval.** For "eigenvector adjusted" centrality, we multiply eigenvector centrality with the primary contribution $\mathbf{U}_1^T\mathbf{y}$. For "seed nonnegative", we only seed if the maximum of the centrality measure is nonnegative, otherwise it is named "seed always". 78

3-7 **Homophily and maximum of contextual centrality when $p\lambda_1 < 1$.** We regress the maximum of contextual centrality on homophily after controlling for $\frac{\overline{\mathbf{y}}}{\sigma(\mathbf{y})}$ and $p\lambda_1$. The y-axis is the OLS coefficients of homophily (with the vertical line as the 95% confidence interval) and the x-axis corresponds to three types of networks. We perform the analysis separately for $\frac{\overline{\mathbf{y}}}{\sigma(\mathbf{y})}$ being larger than, smaller than and equals to zero. 79

3-8 **Relationship between approximated cascade payoff and cascade payoff.** The y-axis and x-axis display the correlation and the spreadability ($p\lambda_1$) respectively. Pearson and Spearman's correlation are shown in blue and orange color respectively. 93

3-9 **Predictive power of contextual centrality without any controls for (a) microfinance and (b) weather insurance.** The y-axis shows the 95% confidence interval of R^2 computed from 1000 bootstrap samples from ordinary least squares regressions controlling for village size. The x-axis shows varying values for $p\lambda_1$, which influences only diffusion centrality and contextual centrality. 96

3-10 **Predictive power of contextual centrality with additional controls for (a) microfinance and (b) weather insurance.** For (a), we use village size, savings, self-help group participation, fraction of general caste members, and the fraction of village that is first-informed as done in [23]. For (b), we use village size, number of first-informed households, and fraction of village that is first-informed. The y-axis shows the 95% confidence interval of R^2 computed from 1000 bootstrap samples from ordinary least squares regressions controlling for village size. The x-axis shows varying values for $p\lambda_1$, which influences only diffusion centrality and contextual centrality. 97

3-11 **Average payoffs with 95% confidence interval when standardized average contribution is -4** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 98

3-12 **Average payoffs with 95% confidence interval when standardized average contribution is -3** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 99

3-13 **Average payoffs with 95% confidence interval when standardized average contribution is -2** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 99

3-14 **Average payoffs with 95% confidence interval when standardized average contribution is -1** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 99

3-15 **Average payoffs with 95% confidence interval when standardized average contribution is 0** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 100

3-16 **Average payoffs with 95% confidence interval when standardized average contribution is 1** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 100

3-17 **Average payoffs with 95% confidence interval when standardized average contribution is 2** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 100

3-18 **Average payoffs with 95% confidence interval when standardized average contribution is 3** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 101

3-19 **Average payoffs with 95% confidence interval when standardized average contribution is 4** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. 101

3-20 **Average cascade payoff for variations of contextual centrality and degree centrality.** 102

3-21 **Average cascade payoff for variations of contextual centrality and diffusion centrality.** 103

3-22 **Average cascade payoff for variations of contextual centrality and katz centrality.** 103

3-23 **Comparison of seeding strategies when $\overline{\mathbf{y}}(\mathbf{U}_1^T\mathbf{y}) < 0$ for (a) continuous and (b) discrete.** 105

4-1 **Motivating example.** Each individual is characterized by hobbies and educational background. The black and green link correspond to professional and social networks, respectively. 111

4-2 **Predicting customer preferences on businesses as a matrix completion task.** 112

4-3 Distributions of average ratings of businesses and users. 117

4-4 **Spatial distributions of businesses and users.** The spatial location of a user is the weighted average location of the businesses she has reviewed. The color code represents the average ratings of businesses and users. 119

4-5 **Relationship between spatial distance, difference in average rating, cosine distance between the business attribute vector, and cosine distance between the check-in time vector.** The middle line, lower and upper boundaries of the box (interquartile range or IQR) correspond to mean, median, 25% and 75% of the data, respectively. The lower and upper whiskers extend maximally 1.5 times of IQR from 25 percentile downwards and 75 percentile upwards, respectively. 120

4-6 **The count of user pairs, difference in average rating, and cosine distance between the user metadata vector, concerning the degrees of separation in the Yelp friendship network.** The middle line, lower and upper boundaries of the box (interquartile range or IQR) correspond to mean, median, 25% and 75% of the data, respectively. The lower and upper whiskers extend maximally 1.5 times of IQR from 25 percentile downwards and 75 percentile upwards, respectively. 121

4-7 **Basic idea behind a graph attention network.** The attention weight α_{ij} represents the relevance of neighbor v_j in updating information on v_i. 125

4-8 **The proposed geometric deep learning architecture for learning latent user representations U and business representations V, via iterations over three layers: a dense layer, an LTSM layer, and a GAT layer.** The predicted $\hat{X}$ is obtained using the final updates of U and V via $\hat{X} = UV^T$. 127

4-9 **Distribution of attention weights.** 131

4-10 **Attention weights for a focal (a) user and (b) business.** Nodes are colored by average rating of the user or business, and the intensity of a link represents the attention weight. 131

4-11 **Regression coefficients with confidence intervals for explaining the leading eigenvector of V^TV.** 133

4-12 **Analysis of business categories of different business clusters in Cleveland Heights.** The relative sizes of the words correspond to the frequencies of the category. 135

4-13 **Analysis of business categories of different business clusters in Urbana.** The relative sizes of the words correspond to the frequencies of the category. 135

5-1 **Performance of the proposed algorithm and baselines in the setting of independent (top) and homophilous (bottom) marginal benefits.** The red triangle, the middle line, lower and upper boundaries of the box (interquartile range or IQR) correspond to mean, median, and 25/75 percentile of the data, respectively. The lower and upper whiskers extend maximally 1.5 times of IQR from 25 percentile downwards and 75 percentile upwards, respectively. 151

5-2 **Performance of Algorithm 3 versus structural properties of the network.** 153

5-3 **Performance (AUC) of Algorithm 2 with respect to $\rho(\beta\mathbf{G})$, θ_1, and θ_2.** 160

5-4 **Performance (AUC) of Algorithm 3 with respect to $\rho(\beta\mathbf{G})$, θ_1, and θ_2.** 161

5-5 **Performance of Algorithm 3 versus number of games (top) and noise intensity in marginal benefits (bottom).** 162

5-6 **Performance of Algorithm 3 versus strength of homophily in the marginal benefits.** 162

5-7 **Performance of Algorithm 2 versus number of games (top), noise intensity in marginal benefits (middle), and structural properties of the network (bottom).** 163

5-8 **Clustering of Swiss cantons based on the political network learned by Algorithm 2 (left) and Algorithm 3 (right).** 164

A-1 **Graphical representation of the Stochastic Block Influence Model (SBIM).** Assume there are two communities, a high socioeconomic status (SES) group (colored in dark blue) and low SES group (colored in dark red), characterized by multi-dimensional sociodemographic features. The two groups have higher intra-class connection probability and lower inter-class connection probabilities. The decision-making of A is jointly influenced by her preferences, as well as her neighbors from the same and different communities. 196

A-2 **Size of each social block.** The y-axis corresponds to the number of individuals in the block, and the x-axis is the corresponding block index. 201

A-3 Adjacency matrix sorted by the inferred block index. The x-axis and y-axis correspond to the indices of individuals. The white and black cells correspond to the existence and the non-existence of edges. We can clearly observe the underlying communities from the network. . . 203

A-4 Interaction matrix and influence matrix. 208

A-5 **Sociodemographic analysis of each social block and the social influence across social blocks.** Each node represents a social block corresponding to the index shown in the previous in Table A.2. The directed links represent the strength of social influence varying from strong negative (blue) to strong positive (red). The color of the node represents a measure of the sociodemographic characteristics within that social block. We display a subset of characteristics, including median age, gender ratio, caste diversity, language diversity, and profession diversity within each block. 209

List of Tables

2.1 **Basic statistics about the mobile phone data set in country A.** 38

2.2 **Basic statistics about the mobile phone data set in the city in country B.** 38

3.1 **Centrality measures defined by $\mathbf{c}_t = \alpha \mathbf{A}\mathbf{c}_{t-1} + \boldsymbol{\beta}$.** 91

4.1 **Summary statistics of the data.** 116

4.2 **Notations** 123

4.3 **Performance comparison using RMSE metric (standard errors in parentheses).** 129

4.4 **Cluster characteristics for Cleveland Heights (C.H.) and Urbana (U.).** 134

5.1 **Performance (R^2) of learning marginal benefits.** 154

5.2 **Performance (AUC) of learning the structure of the social network and the trade network.** 156

A.1 Model and baseline performance 200

A.2 **Block characteristics example. SES is an abbreviation for socioeconomics status.** The majority refers to the largest subset. Disadvantaged caste refers to lower castes, including the castes OBC (Other Backward Class) and Scheduled. Higher education refers to having education levels at PUC (pre-university course) and having a "degree or above" designation. Moderate and lower education levels include all levels below this, where moderate levels have more SSLC (Secondary School Leaving Certificate) levels, and PUC levels and lower levels have mostly primary school education levels. 204

A.3 **Block attributes associated with different types of influence.** Positive and negative influence refers to the type of influence from one block to another block. Self-influence refers to positive influence within the same block. Overlap refers to overlapping categories, such as caste type, profession type, education levels, or languages spoken. 205

Chapter 1

Introduction

Telescope and microscope are two revolutionary technologies in astronomy and biology. They enable human beings to see things that are invisible to naked eyes, either is too small or too far away. The capabilities to collect and analyze massive amounts of data dramatically change these fields. A similar revolution is currently taking place in social science. Digital technologies are similar to telescope and microscope and are named as "socioscope" [161]. They enable researchers to collect information about human behavior that are invisible through lab experiments and surveys. They capture information about apps we use, the places we travel to, people we interact with, and transactions we make. The large-scale, high-resolution, longitudinal, and dynamic information enables us to ask a much broader set of questions on human behaviors, which are not possible through traditional lab experiments and surveys. Yet, the increasing volume, complex structures, and dynamics of behavioral data, as well as the research questions on the complicated human interactions, stretches the limit of conventional methods. The emerging field of data-driven "Computational Social Science" is sparked by the massive amounts of digital records of human behaviors [128]. Before 2000, research on human interactions relied heavily on or one-shot and self-reported data, and data on rational decisions were mostly collected via surveys or lab experiments. Even though lab experiments allow precise control of extraneous variables—making it easier to establish causal relationships and understand the underlying mechanism, the artificial setting makes it difficult to generalize to real

life.

With digital technologies, researchers can perform observational studies with passively collected data sets—such as mobile phone records, credit card transactions, or social media data—on a large-scale (more representative comparing with traditional data collections) population and in a dynamic fashion. New technologies offer at least three advantages. First, the data provides much higher resolution information both at a temporal and spatial scale, including granular information about user behaviors and individuals connections. This enables us to understand the underlying mechanism for the human decision-making process and design interventions accordingly. Second, digital technologies enable us to have a more comprehensive picture of how a 'macro' social network performs, including human behaviors at a city scale for an extended period. Third, digital platforms capture information dynamically, enabling us to understand how society evolves.

1.1 Background

Collective behavior over social networks It is a widely-known phenomenon in social science that collective behavior is not simply adding up individual behavior; because individuals interact. They learn from, strategically interact with, exert peer pressure on each other. At the same time, the social norm emerges. The social network is interesting not only because it influences how decisions are made; but also because of the rich information contained in the interactions. On the one hand, no decision is made in isolation. Specifically, social influence leads to associations in decisions among neighbors and can help organizations to spread products and policy-makers for population management. On the other hand, not merely a medium to connect people, social networks also contain rich information about individuals due to the endogenous network formation and the dependencies in decision-making. The investigation of human behaviors over social networks enables us to understand a wide range of domains, including innovations, competing technologies, cultural fads, social norms, cooperation, social disorder, or financial markets. Hence, how decisions are

made in a network environment has raised interests and has overarching applications in such diverse fields as sociology, economics, health, and political science [14, 23, 63, 44].

A need for new computational tools to deal with the complexities in both human decision-making and the behavioral data structures. The perplexities in human behavior and the increasingly complicated data structures stretch the limits of traditional methods, e.g., data analysis and research designs. There is a need for new computational techniques to understand, predict, and intervene in human decisions, in response to the rich and heterogeneous behavioral information. More recently, technological advances networks facilitate social interactions that otherwise would not take place. These social interactions, therefore, raise many fundamental questions, such as how social influence change short-term decision-making and long-term habits, how do individuals make strategic decisions in a networked environment, how can organizations incentivize behavioral changes by leveraging network effect.This thesis will tackle these questions with new computational methods and large-scale datasets.

1.2 Research questions

My thesis answers questions related to how social influence affects decision-making, how to leverage essential individuals for network-based interventions, how to predict preferences using network information, and how to recover networks based on individuals' actions. To achieve this, I develop tools to integrate individual rational decision-making and machine learning to help with learning problems on social interactions. I will describe the research questions in more detail as follows.

How does social influence spread over social networks? How does the information aggregation process influence the diffusion of social influence?

Several empirical studies have shown that social influence propagates beyond direct neighbors is relatively costless online decision-making settings. Yet, precisely how influence plays a role in costly offline behaviors and spreads through a social network

remains unclear. I leverage the high-resolution mobile phone data and a new behavioral matching framework to study how social influence propagates and affects individual offline behavior. The results show that propagation within the network persists in shaping individual decisions through up to three degrees of separation. To further understand the diffusion of social influence on offline adoption decisions, I propose a Bayesian learning model based on local information aggregation, which better predicts individual adoption behavior than exposure-based contagion models.

How can we design a centrality measure to incorporate both the network connections and the characteristics of individuals nodes? Existing centrality measures study the connectedness of individuals. However, these measures are less helpful in some applications where the objective is to target users who spread positive influence, such as viral marketing or political campaigns. I develop the "contextual centrality" to guide such applications. In particular, contextual centrality evaluates individuals' importance based on network positions and nodal characteristics. It generalizes over existing centrality measures and provides insights on both local and global diffusion. Contextual centrality is shown to perform better in the empirical analysis and simulations on the marketing campaigns for microfinance and weather insurance in rural Indian and Chinese villages. This work provides building blocks for integrating network structures and note features in future network studies.

How to effectively integrate network connections and auxiliary information about individuals to extract informative network connections for the recommendation? Relative to the prior state-of-the-art recommender systems, which employed either nodal characteristics or network structure—but not both—for the recommendation, our approach enables recommendation systems to combine both sources of information to extract useful components of the network and predict individual preferences using data with a complex structure. I test the methodology to Yelp review data and predict customer preferences for restaurants they have not rated, utilizing information on historical ratings, socio-demographics, business characteristics,

check-in information, geographical information, and social networks. The methodology has a wide range of other potential applications, including behavioral predictions and preference inference.

How can we infer the network connections from observed actions when the networks are unavailable? In many social settings, social connections are either unobserved or noisily measured. Individual actions provide information about the underlying interaction structures due to the dependencies of neighbors' actions. I formalize this idea with a linear-quadratic network game. This game is an approximation of all static games with continuous utility functions. I use Nash Equilibrium to approximate users' actions by assuming that rational agents maximize their utilities. I provide conditions under which network structure can be inverted from observed actions, and I perform several empirical applications of the framework.

Chapter 2

Long range social influence in phone communication network

Several empirical works have shown that, in online decision-making settings, social influence propagates beyond direct contacts, mainly due to the exposure effect explained by simple or complex contagion [96, 59, 44, 80, 16]. Yet precisely how influence affects offline behaviors and propagates through a social network, and especially the underlying mechanism that drives such propagation, remain unclear [66, 172, 44]. In this study, we leverage high-resolution mobile phone data sets and a new behavioral matching method based on revealed preference theory, to study how social influence propagates and affects individual off-line behavior. Our results show that propagation within the network persists in shaping individual decision-making to more than three degrees of separation regarding attending an international cultural performance in a European country and visiting a newly opened retail store in a city in North America. To better understand this long range effect of social influence, we propose a Bayesian learning model based on a local learning and information aggregation process, and show that it leads to better prediction of individual adoption behavior compared to exposure-based models. The present study contributes to a theoretical understanding of the diffusion of influence in social networks, which may have significant implications in a variety of practical domains. [1]

[1] This work is joint with Xiaowen Dong, Matias Travizano, Esteban Moro and Alex Pentland.

2.1 Introduction

The effect of social influence on shaping individual decision-making, in various aspects of daily life, has attracted interest from such diverse fields as sociology, marketing, economics, health, and political science [64, 47, 23, 14, 63, 44]. One important motivation of studying social influence is that it may lead to an cascading effect: one's action may influence their direct contacts, who further diffuse the influence through the contact network. Prominent contagion-based theories [96, 59] in social sciences explain the cascading patterns of such diffusion for certain behaviors, such as adopting an app [189, 13], expressing political preferences [44], or sharing a post on social media [125]. These theories model the adoption behaviors as an outcome of exposure to either a single source of information, such as disease spreading [196] and information spreading [19], or multiple sources of information [59], such as registration for health forum [58] and adoption of hashtags [165].

With the increasing popularity of online social networks, many studies have sought to empirically measure the diffusion of social influence on decision-making in virtual space. Notable examples include [44], which uses a 61-million-person online experiment to show that one's political self-expression can be influenced by the friends of their friends; and [80], which uses public goods game in an online experiment to show that behavioral contagion reaches up to three degrees of separation in a social network. Moreover, [16] conducted a randomized experiment and shows that integrating viral features into commercial applications hosted on Facebook increases the total adoption, due to passive broadcasting messages spread through the Facebook network.

Comparing to the online setting, most studies of offline behavioral contagion, such as that of exercising activities [15], voting behavior [44], and adoption of microfinance [25], is restricted to direct neighborhood in the social network. On the one hand, offline behaviors are associated with a higher cost of communication and decision-making; hence, their diffusion is likely to have a broader socioeconomic impact. On the other hand, this presents significant challenges due to the difficulty in getting quality data for studying large-scale offline behaviors as well as that in measuring the effect of

social influence in such settings. The theory of "three degrees of influence" proposed by [64] is based on small-scale offline experiments focusing on smoking behavior, happiness, and habits that led to obesity, and found that they propagate within a social network up to three degrees of separation. However, the strategy in identifying social influence in their work has raised some criticism [172, 66], and the size and scale of the experiments are relatively limited.

To understand how influence on offline behaviors propagates in a large-scale social network, we leverage two high-resolution data sets of mobile phone communication records from the country A and the one city in Country B, to construct offline communication networks and study the effect of social influence on two types of offline adoption behaviors: i) attending an international cultural performance, and ii) visiting a newly opened retail store. We focus on the effect beyond direct contact or immediate neighbors in the communication network. To control for potential confounding factors such as homophily, we propose a novel matching framework to mimic the random assignment of treatment, conditioning on personal preference that is revealed by historical mobility patterns [133]. This framework allows us to study the propagation of influence in large-scale settings, thereby overcoming the difficulty of applying randomized controlled trials (RCTs) in such scenarios [28]. Our results show that, within the communication network, the effect of social influence decays from the initial adopters' direct contacts but, surprisingly, persists to more than three degrees of separation in both cases. This result suggests that one's decision to adopt and communicate with others may impact individuals in the network far beyond one's direct contacts. Indeed, this pattern of social influence resembles the physical phenomenon of ripples expanding across the water when an object is dropped into it.

Traditional contagion-based models bear attractive mathematical properties and achieve good performance in predicting adoption behavior that is dominated by exposure. Going beyond simple exposure-based models, we are interested in the mechanism behind decision-making where rational individuals maximize their utilities by learning and aggregating information from neighbors in a social network [204]. More specifically, we propose a Bayesian learning model in which individuals dynamically

update their posterior beliefs towards the adoption behavior and make decisions on adoption based on the local information they collect from immediate neighbors. We further show the link between the proposed model and the empirical decaying pattern of the effect of social influence. In a task of predicting future adoption, the proposed model outperforms other state-of-the-art contagion models, including , the independent cascade model [116, 119], the threshold model [95, 195] and the structural econometric model [23], thereby suggesting the importance of incorporating the local information aggregation process into network-based Bayesian learning for predicting individual decision-making.

We make several contributions to the existing research. First, we leverage two large-scale mobile phone data sets to show that in a dynamic phone communication network the effect of social influence on offline decision-making can extend to more than three degrees of separation. This demonstrates the hidden impact of one's decision on that of others beyond one's immediate circle, which may be used for designing network-based intervention and marketing campaigns. Second, our research design can be generalized to study social influence in a wide range of online and offline settings with rich behavioral information. On the one hand, thanks to the availability of mobile phone data (i.e., call detail records) in almost every country in the world, such study becomes possible even when running large-scale experiments might be difficult due to resource constraint. On the other hand, we highlight the effectiveness of behavioral data in revealing users' preferences and other characteristics such as socio-demographics. Behavioral information are not only more commonly-collected on digital platforms, but can also be studied in a dynamical fashion, hence capturing the potential change in users' preferences and tastes. Third, we develop a Bayesian learning model which extends existing contagion models with a dynamic local information aggregation process. This extension accounts for individual heterogeneity and can capture both positive and negative influence signals. The improvement in the prediction performance achieved by the proposed Bayesian learning model demonstrates that the propagation of social influence is more complex in offline than online settings, which calls for the need for a decision-making mechanism that goes

beyond simple contagion or exposure-based models.

The rest of this section of my thesis is organized as follows. Section 4.2 presents a review of the literature. Section 2.3 describes the data used in this chapter, and proposes the behavioral matching strategy. Section A.4 presents the empirical results with robustness analysis, and section A.3 develops a Bayesian learning model with a local information aggregation process. Section 2.6 highlights managerial implications and Section 2.7 present discussion of the results of the study.

2.2 Literature review

2.2.1 Contagion models

There are two prominent theories in the literature for explaining the propagation of online behaviors [189, 13, 44, 125], i.e., simple contagion and complex contagion. Simple contagion theory assumes that individuals will adopt the behavior as long as they have been exposed to the information [96], which is a sensible model for disease and information spreading. Complex contagion theory, on the other hand, requires multiple sources of information (i.e., social reinforcement) to trigger the adoption [59]. Studies have shown that complex contagion explains contagion behaviors such as registration for health forums [58].

While these exposure-based models have been shown to explain the farther diffusion in the virtual space where decision-making is relatively effortless, it is not clear whether they can explain offline decisions that are associated with higher time and socio-economic cost. More importantly, these models cannot capture the potentially negative effect of social influence, i.e., the adoption decision of one's neighbors might decrease, rather than increase, the likelihood of one's adoption decision. Traditional contagion models, despite their simplicity, do not take into account heterogeneity in individual preferences which may lead to such negative influence.

2.2.2 Observational learning and word-of-mouth effect

There has been a broad interest in marketing literature to study the mechanism of word-of-mouth (WOM) effect [209, 150] and observational learning (OL) [101, 207] on product adoptions. In WOM, consumers infer product information directly from others' opinions, while in OL, consumers infer information about products from others' previous actions indirectly [8]. [61] uses a natural experiment to measure the effects of both WOM and OL on product sales on Amazon. It has been shown that OL is likely to lead to an information cascade such that all subsequent observers would share similar beliefs about the underlying parameter they try to estimate in making adoption decision [26]. For example, [75] demonstrates that informational cascade leads to herding behavior in online software adoption, whereas [208] examines whether such herding behavioral is rational or not. Furthermore, [175] and [164] study restaurant discovery using OL from friends and strangers.

In the WOM and OL literature, individuals follow or un-follow the behaviors of neighbors without trying to estimate the underlying characteristics of the product in making adoption decisions. Therefore, they inherently assume that individuals are homogeneous in tastes and preferences, which might be unrealistic [204]. The proposed Bayesian learning model differs from these models by enabling an individual to learn about (1) how similar her friends are to herself in terms of preferences and (2) whether her friends are positive about the product or not. As we shall see, this naturally allows the proposed model to capture both the positive and negative effect of social influence in adoption.

2.3 Behavioral matching framework

Adoption behaviors in a social network are widely seen as resulting from two factors: similarities among friends (i.e., homophily) and contagion driven by social influence [136, 14, 13, 44, 189]. To quantify the effect of social influence while controlling for the upward estimation bias caused by homophily, traditional studies in the literature mostly rely on socio-demographic information as covariates for characterising individuals

[17, 14, 13, 44]. However, such information is not always available and cannot adapt to changes in individual tastes and preferences.

To this end, we propose a novel framework where we make use of mobile phone data records to compute the revealed preferences of individuals. In the economics literature, the theory of revealed preference is used to estimate consumers' preferences based on their observed buying decisions [170, 200]. Similarly, we infer individual preferences using their leisure activities over the weekend, which are based on observed choices of destinations, i.e., the frequency with which they visit different places [133, 112, 141]. In particular, this information serves as proxies for activities individuals perform in their spare time, and can approximate their income from their home and work locations. A similar study controlling for behavioral covariates shows that the average treatment effect estimated by controlling for high-dimensional behavioral covariates reduces the estimation bias by 97% compared with a random experiment on Facebook [76].

2.3.1 Setting

We consider two large-scale mobile phone data sets, one collected in country A and another in one city in country B, that include individual phone communication records as well as the location of the cell tower to which each call was connected. In the case of country A, the data set covers a period of seven months from January 2016 to July 2016. In the case of country B, the data set comprises a period of twelve months from October 2015 to September 2016. Table 2.1 and Table 2.2 show basic statistics of the mobile phone data sets used in this study. We consider two offline adoption behaviors, i.e., attending an international cultural performance in country A and visiting a newly opened retail store in country B. For notational convenience, we consider both the performance and the store as "products", and the attendees and visitors as "adopters".

We discretize our overall data sets into different observation periods, which can overlap depending on the type of adoption behavior being considered. We define each observation period $T = [s, s + \tau]$, where $T \in \Psi$, $s \in \mathcal{S}$, and τ is the length of each period. Here, Ψ is the set of all observation periods, and $\mathcal{S}$ is the set of the starting

Table 2.1: **Basic statistics about the mobile phone data set in country A.**

	mean	**standard deviation**
number of months	7.0	/
average number of calls per month	5893779.4	386793.9
median number of calls per person	3.0	0.6
total number of calls per person per month	27.9	4.8
median number of friends per person per month	1.9	0.4
average number of friends per month	5.6	0.7
number of individuals per month	217832.3	45782.5

Table 2.2: **Basic statistics about the mobile phone data set in the city in country B.**

	mean	**standard deviation**
number of months	12.0	/
average number of calls per month	7634397.4	1089728.8
median number of calls per person	43.6	30.6
total number of calls per person per month	91.7	27.3
median number of friends per person per month	18.2	2.2
average number of friends per month	11.2	2.2
number of individuals per month	646163.2	37403.7

time instances of each period. In the case of country A, for each performance day we choose the observational period T to be a period of $\tau = 24$ hours, starting with the beginning of the performance. In the case of Merida, for each day within three months after the store was opened, we choose T to be a period of $\tau = 72$ hours, starting with the beginning of the day. The motivation behind a different τ for the two data sets is as follows. The cultural performance took place only on the weekdays in a given month, therefore individuals needed to decide in a reasonably short period. In comparison, the influence on the decision to visit the store may take longer to appear, both because people do not go to stores every day and because they know that the store would remain open for an extended period.

We introduce three key concepts in this section of my thesis. First, we define $\mathcal{D}_T$ as the set of **initial adopters** for the observation period T. We identify people as initial adopters if they were connected to a cell tower close to the performance venue or the store location during a time interval at the beginning of the period T. In the case of country A, such an interval is defined as the time window of the performance

(with a buffer time of ± 30 minutes). In the case of Merida, the interval is defined as the first day in T. Second, we construct an **information cascade** as a directed graph $\mathcal{C}_T = (\mathcal{I}_T, \mathcal{E}_T)$, where $\mathcal{I}_T = \{1, 2, 3, ..., n\}$ is a set of n individuals who have at least one cell phone activity in T, and $\mathcal{E}_T = \{(i, j)\}$ is a collection of ordered node pairs (i, j) conditioned on that i possesses information about the product when the communication with j took place and that i spreads the information to j. Next, we define a path of length $K-1$ between individual i and j in $\mathcal{C}_T$ as a sequence of distinct individuals, $i_1, ..., i_k$, such that $i_1 = i$, $i_k = j$, and $(i_k, i_{k+1}) \in \mathcal{E}_T$ for $k \in \{1, ..., K-1\}$. The social distance from individual i to individual j, $\text{sd}(i, j)$, is defined as either the length of the shortest path from i to j in the cascade $\mathcal{C}_T$ if such a path exists, or $+\infty$ otherwise. This allows us to define the third concept, **hop index** for individual i, as the minimum length of the shortest paths from i to all $j \in \mathcal{D}_T$:

$$h_i = \min\{\text{sd}(i, j) |\ \forall j \in \mathcal{D}_T\}, \text{for } i \in \mathcal{I}_T. \tag{2.1}$$

Therefore, an individual i of hop index h is h-degree of separation from the closest adopter in $\mathcal{D}_T$. An illustration of these three concepts are shown in Figure 2-1.

For the cultural performance in country A, 19 observation periods generated 16,043 adopters in total. In both cases, we construct the information cascade for each T, as shown in Figure 2-1 and compute the hop index for each individual in $\mathcal{I}_T$. The adoption likelihood of each hop are shown in Figure 2-11 of section 2.7. The information cascades cover 161,857 individuals (who are in various treatment groups with different h), with another 71,337 disconnected from the information cascades (who are in the control group). For the store in Merida, 123 observation periods in the three months after the store opened generated 4736 adopters in total. There are 86,413 and 106,340 individuals involved in and disconnected from the information cascades, respectively.

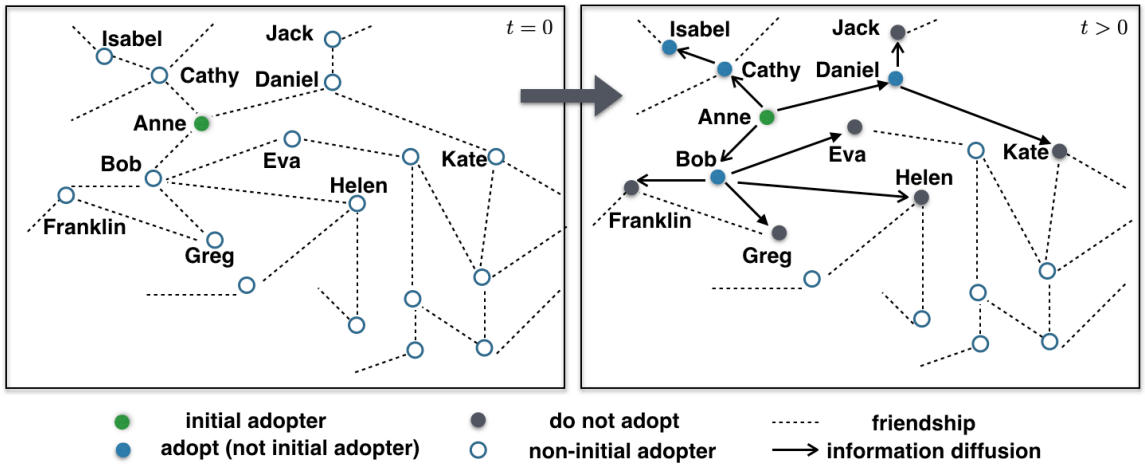
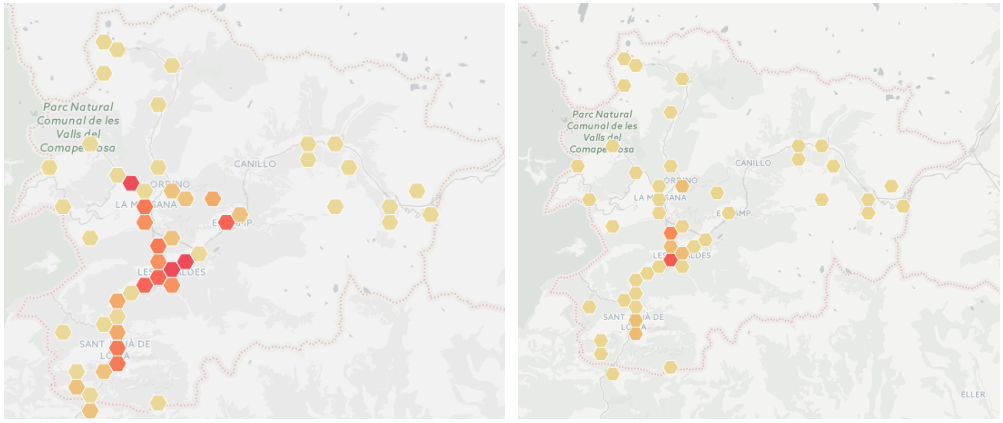


Figure 2-1: **An illustration of initial adopters, information cascade, and the hop indexes.** In the information cascade $\mathcal{C}_T$ shown in the figure, within an observation period T, the initial adopter Anne (colored green) passes information to her neighbors, Bob, Eva, Cathy, and Daniel, who after receiving information from Anne continue to pass the information onwards. Labeled with hop index one, they further diffuse the information to Franklin, Greg, Helen, Isabel, Jack, and Kate, who are then on hop index two. The process continues until the end of the observation period. Among the people who receive information, Bob, Isabel, and Daniel (colored blue) decided to adopt the behavior, while others (colored grey) decided not to.

2.3.2 Matching framework

We use propensity score matching to yield the estimate of social influence by conditioning on individual preference revealed by mobility patterns [120, 166]. Specifically, we consider an individual-destination matrix $\mathbf{M}$ where the j-th row and i-th column correspond to the j-th destination (location of the j-th cell tower) and i-th individual, respectively, and $\mathbf{M}_{ji}$ represent the number of times that individual i has visited destination j during a period prior to the observation periods (around six months in both cases). We then project $\mathbf{M}$ onto a subspace spanned by the top eigenvectors of its covariance matrix to obtain an eigen-preference matrix in which the i-th column, $\boldsymbol{x}_i$, represents the preferences of individual i. Specifically, we choose 14 eigenvectors for the adoption of attending the cultural performance and 44 for the adoption of visiting the retail store, such that they explain at least 80% of the variance of the covariance matrix in each case.

As illustrative examples, we show how individual preferences are revealed by



(a) an explorative pattern (b) an exploitative pattern

Figure 2-2: **Two types of mobility frequency patterns during weekends, revealing different individual preferences: (a) an explorative pattern; (b) an exploitative pattern.** The intensity of the color represents the normalized visitation frequency, i.e., the darker red color corresponds to a more frequent visit.

mobility frequency patterns in Figure 2-2, where the color intensity represents the normalized visitation frequency. Figure 2-2(a) describes the mobility history of an individual with diverse activity patterns, i.e., a person who explores various parts of the country during the weekends (explorative pattern), while Figure 2-2(b) describes the mobility history of an individual who spends most of her weekends in crowded shopping districts (exploitative pattern).

For the information cascade $\mathcal{C}_T$ constructed for each observation period T, we define a treatment group in hop h as the group of individuals with a finite hop index h (and thus of social distance h from the closest adopters in $\mathcal{D}_T$) and a control group in which individuals have an infinite hop index (and thus are not connected to any adopter in $\mathcal{D}_T$). In each period, T, then, there are multiple treatment groups, one for each finite hop index. For each treatment group, every individual is then matched to one in the control group using propensity score matching [120, 166], where the propensity score of being treated in hop h is defined as the conditional probability of being connected to the initial adoption via h hops that is estimated using individuals' preferences via logistic regression. The propensity score matching operates under the conditional unconfoundedness assumption, the adoption behaviors are independent of

the exposure, and that all individuals have a positive conditional probability of being exposed to the information or otherwise. Therefore conditional unconfoundedness implies that exposure to social influence is also unconfounded conditional on propensity score [120, 166].

Under the proposed matching framework, the difference in adoption rate between each treatment group and the control group is the difference in adoption likelihood due to social influence for that particular treatment. For example, the difference in the adoption likelihood for the treatment group in hop h, ΔA_h, is computed as,

$$\Delta A_h = \frac{1}{|\Psi|} \sum_{T \in \Psi} \frac{1}{|\mathcal{M}_T|} \sum_{m=1}^{|\mathcal{M}_T|} (z_m^h - z_m^c), \tag{2.2}$$

where z_m^h and z_m^c are the adoption decisions of the individuals in the m-th matched pair from the treatment group on h and the control group, respectively, $|\mathcal{M}_T|$ is the cardinality of the set $\mathcal{M}_T$ that contains all matched pairs in period T, and $|\Psi|$ is the total number of observation periods. The adoption rate of the control group is denoted as A_0. The difference in the adoption likelihood between the two groups due to social influence thus establishes an upper bound[2] of the extent to which social influence, rather than homophily, explains the adoption behavior [14].

2.4 Long range of social influence

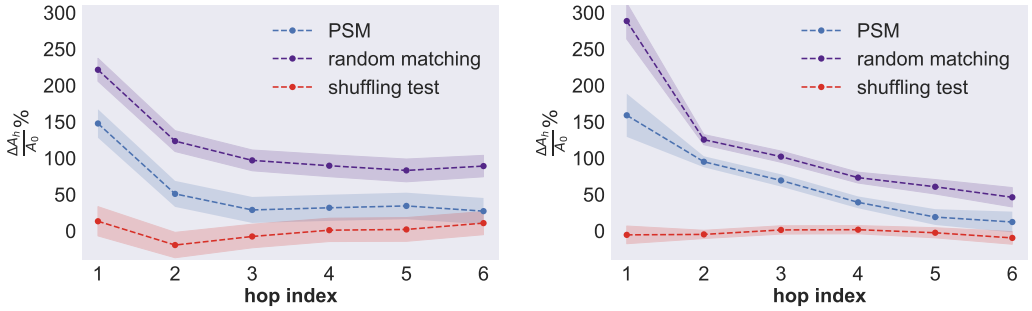
We apply the behavioral matching framework mentioned section 2.3 to estimate the treatment effect of social influence on the likelihood of adoption for the two behaviors under consideration. In Figure 2-3 (a) and (b), the purple dashed line shows the the difference in the adoption likelihood of the treatment group and the control group due to social influence (y-axis) concerning different hop indexes (x-axis). For both

[2]This is mainly due to the difficulty in controlling for unobserved confounding variables using matching-based methods in observational studies. For country A, to partly address the issue that tourists may travel together, we remove individual pairs who are potentially on the same trip to country A. This can be inferred based on whether individuals have stayed at the same hotel on the same nights. For country B, however, it is difficult to verify whether people have received advertisements about the new store via mails, TV, or online sources.

attending a performance and visiting a store, social influence increases the likelihood of future adoption. The effect of social influence is particularly strong for individuals who had direct phone communication with the past adopters, with an increase of 148% (attending a performance) and 169% (visiting a store) in the likelihood of future adoption. To study whether behavioral patterns omit some information contained in socio-demographics, we compare the matching estimates with merely behavioral covariates, as well as supplementing that with socio-demographics in the case of visiting the retail store. By comparing the point estimates and confidence intervals in Figure 2-4, we see that the estimates by matching with and without socio-demographic are almost the same except for a slight difference in the first hop, which indicates that socio-demographics provide a subset of information embedded in behavioral covariates.

More interestingly, we observe the long range effect of social influence over the communication network, originating from the initial adopters' direct contacts and expanding over longer social distances in the information cascades. Specifically, the difference in adoption likelihood of the treatment group and the control group due to social influence shows a decaying pattern from hop one onwards but persists up to more than three degrees of separation in both cases. This rather surprising result suggests that initial adopters' communication may have a hidden impact on the decision-making of individuals far beyond the immediate contact circle. The comparison between the the difference in adoption likelihood for the treatment and the control group in the two cases also seems to imply a difference in virality between the two adoption behaviors.

We perform some robustness checks on the empirical results. First, to mimic the random assignment of treatment in a controlled experiment, we need to ensure that individual pairs in the treatment and the control groups are sufficiently similar. we first evaluate whether there is sufficient overlap between the individual pairs in the treatment and the control group. In other words, the covariates – the preference vectors $\boldsymbol{x}_j$ in our case – must be balanced between the matched pairs to remove the confounding effects. To this end, we use the standardized mean difference (SMD) to evaluate whether the covariates in the treatment and control groups demonstrate sufficient overlap [68]. The SMD is calculated as the difference in means in the unit of



(a) attending cultural performance

(b) visiting retail store

Figure 2-3: **Percentage improvement in adoption rate relative to the control group due to social influence via phone communication network (ΔA_h)**: (a) attending cultural performance; (b) visiting the retail store. The y-axis is the difference in adoption likelihood of the two groups, and the x-axis is the hop index. The purple, blue, and red dashed lines show the estimated effect of social influence using PSM, random matching, and PSM after a shuffling test, respectively. The shaded regions correspond to the 5% and 95% confidence intervals from bootstrap sampling. The higher and lower end of the vertical line indicates the 5% and 95% interval.

pooled standard deviation as follows:

$$\text{SMD} = \frac{\overline{\boldsymbol{x}}_{j,h} - \overline{\boldsymbol{x}}_{j,c}}{\sqrt{(\sigma_{j,h}^2 + \sigma_{j,c}^2)/2}}, \tag{2.3}$$

where $\overline{\boldsymbol{x}}_{j,h}$ and $\overline{\boldsymbol{x}}_{j,c}$ are the means of the covariates $\boldsymbol{x}_j$ for the treatment group on hop h and the control group, respectively, and $\sigma_{i,h}$ and $\sigma_{i,c}$ are the standard deviations of covariates $\boldsymbol{x}_j$ for the treatment group on hop h and the control group, respectively. As a rule of thumb, SMD of less than 0.1 for a particular variable demonstrates sufficient overlap between the treatment and control groups for that variable. All the variables we choose in Figure 2-5 and 2-6 pass this robustness check.

Second, we show as the blue dashed line in Figure 2-3 the estimated influence without controlling for homophily where, instead of the PSM strategy, a member of the treatment group is randomly matched to another in the control group. In both cases, we observe an overestimation of the effect of social influence by about 100% with random matching, which is consistent with the finding in a previous study on the adoption of an online application [14]. To further validate results on the effect of

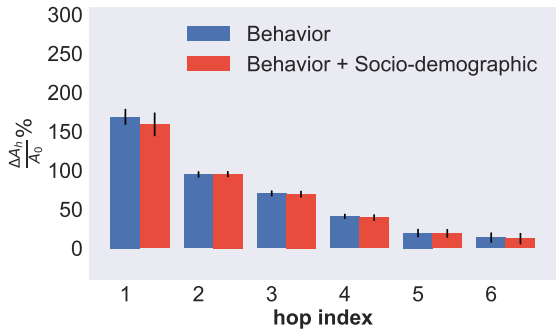


Figure 2-4: **Matching on behavioral covariates, and on both behavioral covariates and socio-demographics, for the adoption behavior of visiting the retail store.** The y-axis is the difference in adoption likelihood of the two groups, and the x-axis is the hop index. The bar plot and the vertical lines correspond to the mean, 5%, and 95% confidence intervals, respectively. The blue and red bars correspond to behavioral matching, and behavioral + socio-demographics matching.

social influence, we test the "random shuffling" strategy proposed by [9] to exclude the effect of homophily or unobserved confounding variables that may induce statistical correlation between the actions of friends and therefore generate the observed decaying patterns. To this end, we randomly assign individuals to the control and treatment groups, with a randomized hop index for individuals assigned to the latter. We then compute the difference in the adoption likelihood due to social influence using the PSM strategy, and the results are shown as the red dashed line in Figure 2-3. Both the increased likelihood of adoption and the decaying patterns mostly disappear, which verifies that the observed patterns are not likely to be driven merely by the effect of homophily or unobserved confounding variables.

Finally, we also use a post-Lasso estimation method to estimate the coefficients for treatments with a data-driven penalty [33]. We use post-Lasso logistic regression to estimate the difference in adoption likelihood of the two groups, which applies ordinary least squares to the model selected by Lasso [32, 33]. The results are shown in Figure 2-7. The estimates via the post-lasso logistic regression, as shown in the blue bar plots, also display a long range effect of social influence and penetrates deep into the communication network, which demonstrates the robustness of our results.

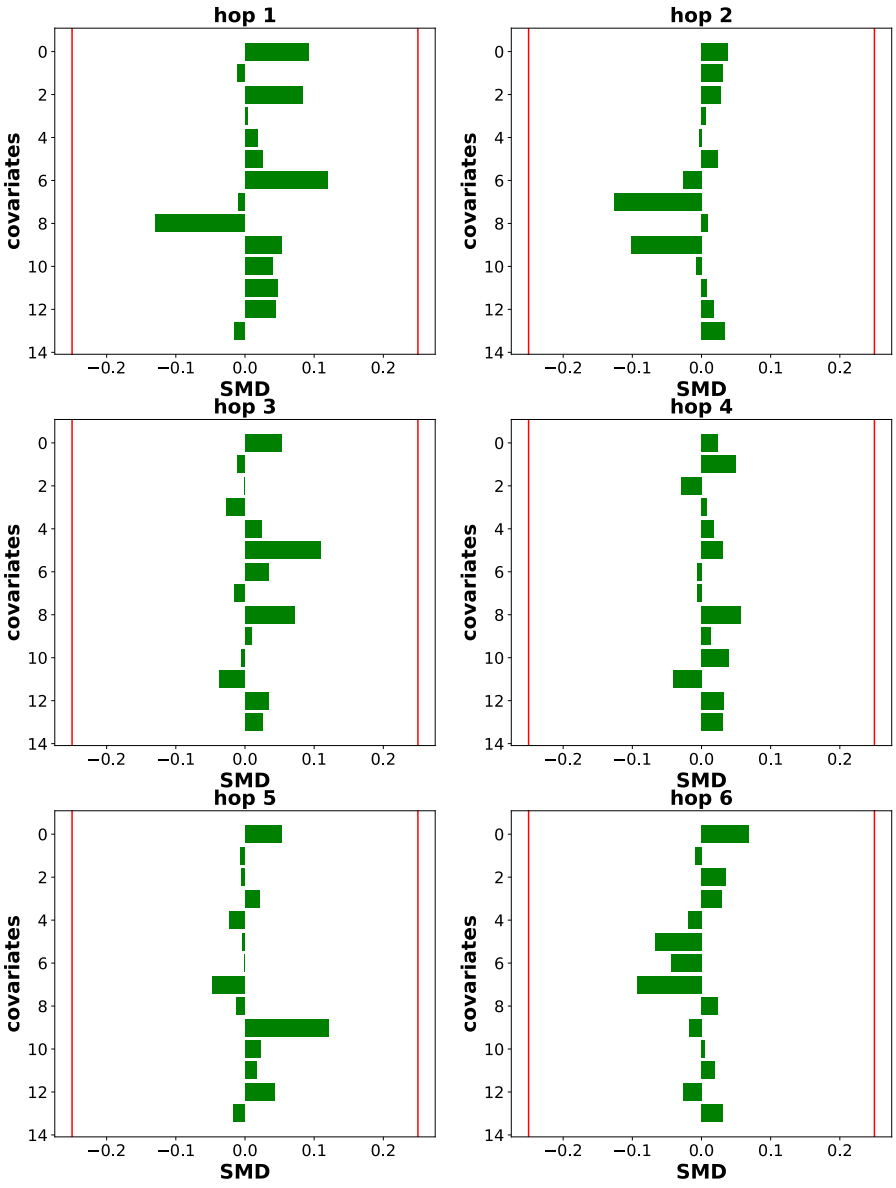


Figure 2-5: **SMD for the matching between the control group and the different treatment groups (different hop indexes h), in the case of attending cultural performance.**

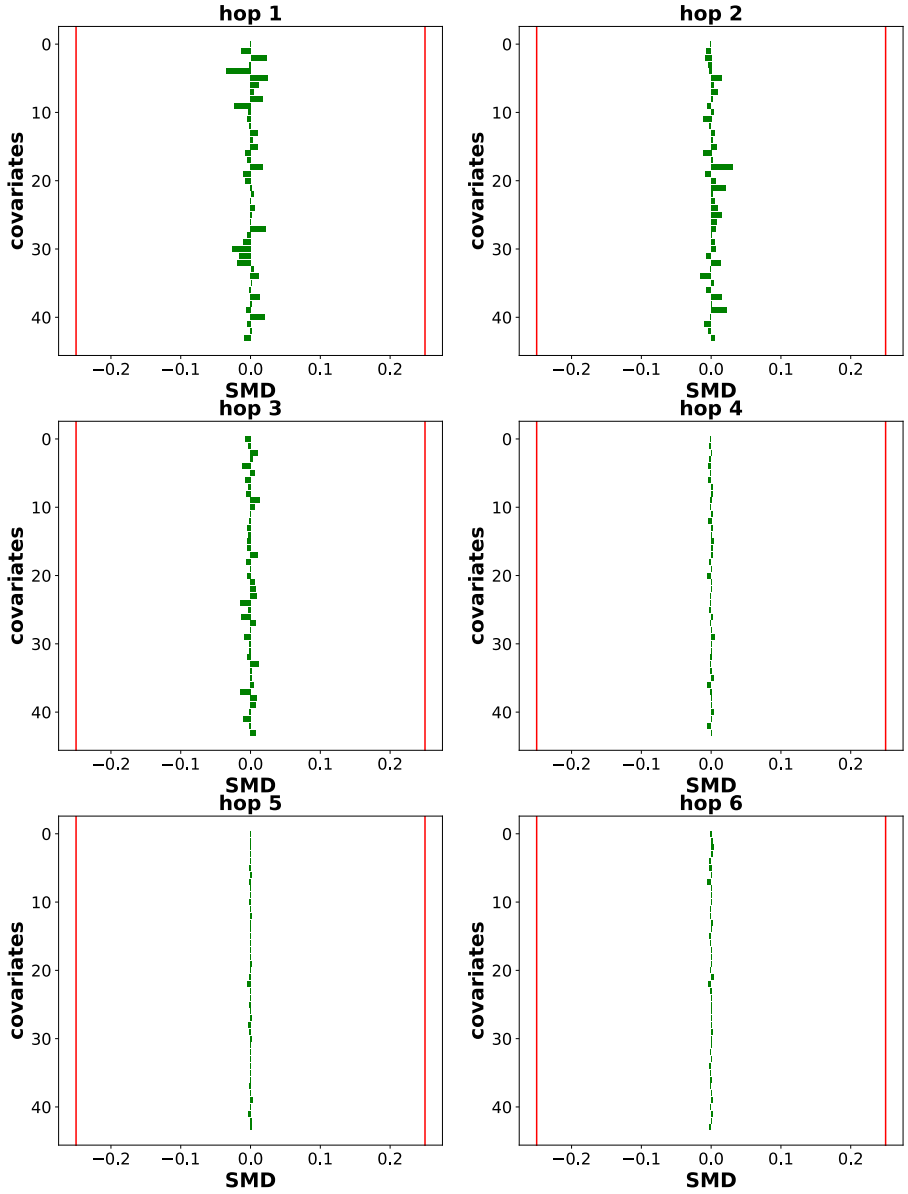
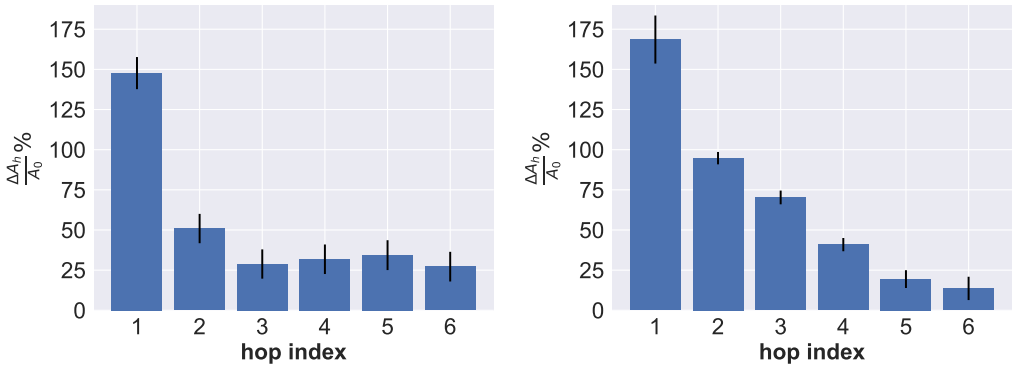


Figure 2-6: **SMD for the matching between the control group and the different treatment groups (different hop indexes h), in the case of visiting retail store.**



(a) attending cultural performance

(b) visiting retail store.

Figure 2-7: **Percentage improvement in adoption rate relative to the control group due to social influence estimated by post-Lasso logistic regression for different treatment groups (ΔA_h)**, in the case of a) attending cultural performance and b) visiting the retail store. The vertical bars cover 5% and 95% confidence intervals.

2.5 Bayesian learning model and results

Motivated by these empirical results, we are interested in investigating the underlying mechanism that leads to the global decaying pattern of social influence's effect to better understand how individuals make adoption decisions.

2.5.1 Bayesian learning model

We build upon the literature of Bayesian learning [3, 4, 140, 149] and propose a framework under which individuals dynamically aggregate local information by sequential communication with their neighbors (defined as people who have called one another within the observation period) to dynamically estimate the product's latent characteristics. However, rather than focusing on observing neighbors and learning from their actions in a hypothetical setting, we explicitly model the dynamic learning process of a high-dimensional latent vector that captures posterior belief about the product based on the preferences and evaluations of immediate neighbors.

There are two critical factors in our model that determine the local decision-making process of an individual i: i) her perception ($\boldsymbol{w}_i \in \mathbb{R}^d$) of the product at each time step t, which is constantly being updated after interaction with neighbors (i.e., the posterior

belief), and ii) her preference ($\boldsymbol{x}_i \in \mathbb{R}^d$), which is assumed to remain unchanged in the relatively short decision-making window (i.e., the observation period). We further make the reasonable assumption that neighbors know each others' preferences ($\mathbf{x}_i$). Let $P_t(\boldsymbol{w}_i)$ and $\boldsymbol{\mu}_{i,t}$ represent the probability density function and expectation of $\boldsymbol{w}_i$ at time instance t, respectively. We use the inner product $\langle \boldsymbol{x}_i, \boldsymbol{\mu}_{i,t} \rangle$ to measure the similarity between $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ evaluated at time instance t. Here we denote $\boldsymbol{\mu}_{i,t}$ as the posterior mean, i.e., mean of $P_t(\boldsymbol{w}_i | y_{j,t}, \mathbf{x}_j)$, which is the perception of the product formed by i at time t. The larger the $\langle \boldsymbol{x}_i, \boldsymbol{\mu}_{i,t} \rangle$, the higher i's evaluation of the product (i.e., i's perceived utility by adopting the behavior). Then, i's evaluation of the product at time instance t, denoted by $y_{i,t} \in \mathbb{R}$, is expressed as:

$$y_{i,t} = \langle \boldsymbol{x}_i, \boldsymbol{\mu}_{i,t} \rangle + \epsilon_{i,t}, \tag{2.4}$$

where $\epsilon_{i,t}$ captures the unobserved variables that might influence i's decision-making at time instance t. We assume that $\epsilon_{i,t}$ follows a zero-mean Gaussian distribution: $\epsilon_{i,t} \sim \mathcal{N}(0, \beta^{-1})$, where β is the precision of $\epsilon_{i,t}$.

Although our model can be generalized to any distribution where the parameters may be estimated using variational inference [40], we assume for the sake of simplicity that the prior and the posterior of $\boldsymbol{w}_i$ follow Gaussian distributions:

$$P_0(\boldsymbol{w}_i) = \mathcal{N}(\boldsymbol{w}_i | \boldsymbol{\mu}_{i,0}, \boldsymbol{\Sigma}_{i,0}), \tag{2.5}$$

where $\boldsymbol{\mu}_{i,0}$ and $\boldsymbol{\Sigma}_{i,0}$ are our priors of the mean and covariance of the product characteristics $\boldsymbol{w}_i$ at time step 0, respectively. We further assume that the likelihood also follows a Gaussian distribution, where for $t \geq 1$ we have:

$$\mathcal{L}_t(\langle \boldsymbol{x}_j, \boldsymbol{\mu}_{i,t} \rangle | y_{j,t}) = \mathcal{N}(y_{j,t} | \langle \boldsymbol{x}_j, \boldsymbol{\mu}_{i,t-1} \rangle, \epsilon_{i,t}). \tag{2.6}$$

The information i receives at time t consists of the preference of j ($\boldsymbol{x}_j$) as well as j's evaluation of the product ($y_{j,t}$), for $i, j \in \mathcal{I}_T, (i,j) \in \mathcal{E}_T$. In particular, i updates her estimation of $P_t(\boldsymbol{w}_i)$ using the Bayes' rule, which is considered as the conditional

probability of $\boldsymbol{w}_i$ given i's information at time step t. Particularly,

$$\begin{aligned} P_t(\boldsymbol{w}_i|y_{j,t}, \mathbf{x}_j) &\propto P_{t-1}(\boldsymbol{w}_i)\mathcal{L}_t(\langle \boldsymbol{x}_j, \boldsymbol{w}_i\rangle|y_{j,t}) \\ &= \mathcal{N}(\boldsymbol{w}_i|\boldsymbol{\mu}_{i,t-1}, \boldsymbol{\Sigma}_{i,t-1}) \cdot \mathcal{N}(y_{j,t}|\langle \boldsymbol{x}_j, \boldsymbol{w}_i\rangle_t, \epsilon_{i,t}) \\ &= \mathcal{N}(\boldsymbol{w}_i|\boldsymbol{\mu}_{i,t}, \boldsymbol{\Sigma}_{i,t}), \end{aligned} \tag{2.7}$$

where i's estimation of the product characteristics at time step $t-1$, $P_{t-1}(\boldsymbol{w}_i)$, is considered as a prior distribution at time t. Correspondingly, $\boldsymbol{\mu}_{i,t}$ and $\boldsymbol{\Sigma}_{i,t}$ are the posterior mean and posterior covariance of $\boldsymbol{w}_i$ at time step $t \geq 1$. Following [39] and [38], we can then derive estimates for the posterior mean and covariance as follows:

$$\begin{aligned} \boldsymbol{\Sigma}_{i,t}^{-1} &= \boldsymbol{\Sigma}_{i,t-1}^{-1} + \beta \boldsymbol{x}_j^T \boldsymbol{x}_j, \\ \boldsymbol{\mu}_{i,t} &= \boldsymbol{\Sigma}_{i,t}(\boldsymbol{\Sigma}_{i,t-1}^{-1}\boldsymbol{\mu}_{i,t-1} + \beta \boldsymbol{x}_j^T y_{j,t}) \end{aligned} \tag{2.8}$$

We illustrate the updating process using Greg's dynamic information aggregation and decision-making as shown in Figure 2-8, which can be summarized in the following steps:

1. For an observation period T, an initial adopter j in $\mathcal{D}_T$ communicates with her neighbors about the product with probability p and express her evaluation $y_{j,t}$ at time instance t.

2. A particular neighbor i, who has received the information updates her posterior on the perception of the characteristics of the product, based on the preferences and product evaluations of the initial adopter (e.g., j) from whom the neighbor (e.g., i) receives the information.

3. Based on the updated posterior, i forms her posterior evaluation ($y_{i,t}$) and makes a decision on adopting the product ($a_{i,t}$). The probability that i adopts the product at time step t is defined as:

$$\log(\frac{a_{i,t}}{1-a_{i,t}}) = y_{i,t}. \tag{2.9}$$

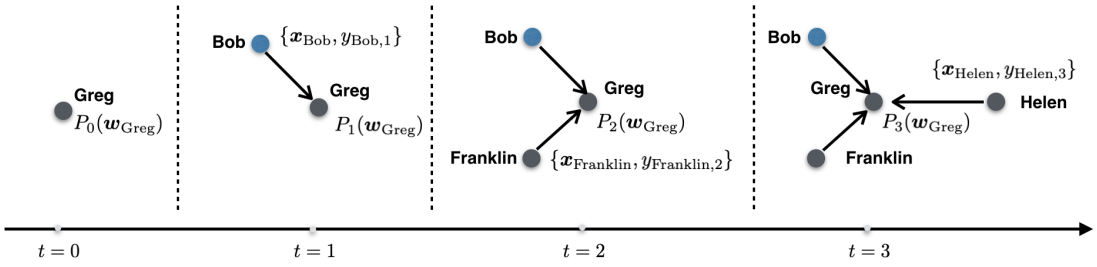


Figure 2-8: **Decision-making process for Greg, according to the proposed Bayesian learning model.** At the time instance $t = 0$, Greg forms a prior understanding of the product. At $t = 1$, Bob told Greg about his evaluation of the product. Knowing Bob's general preference, Greg then updated his perception ($P_1(\boldsymbol{w}_{\text{Greg}})$). The same updating process happens after observing the preferences and the evaluations of Franklin and Helen afterward. With this illustration, we show how Greg updates his perception about the product by dynamically aggregating local information from his neighbors who communicated with him.

4. In the subsequent iteration of the information cascade, i communicates with her neighbors about the product with different probabilities. If i decides to adopt, she diffuses the information with probability p_a; otherwise, she diffuses the information with probability p_n. If she diffuses the information to her neighbors, the neighbors then repeat steps 2 to 4.

Our Bayesian learning model captures two important micro-processes happening when social influence spread that has been overlooked by existing contagion models. In particular, the evaluation of the information receiver is influenced by two factors: 1) the similarity between the spreader and the receiver; 2) the evaluation of the spreader. Our method naturally accommodates negative social influence in the following way, which fills the gap in existing contagion models [116, 59, 23]. Specifically, social influence is negative with the following two mechanisms: if the spreader evaluate the product negatively and the similarity between the spreader and the receiver is high; or if the spreader evaluates the product positively, and the similarity between the spreader and the receiver is low.

2.5.2 Comparisons between the Bayesian learning model with existing models

We briefly discuss the differences between our Bayesian learning model and existing models proposed in the literature, including both contagion models and Bayesian learning models.

First, the widely-used independent cascade model [116, 119] and threshold model [95, 195] both provide a mechanical way to explain the contagion process with a single parameter, i.e., a diffusion rate for the former and a decision threshold for the latter. However, unlike in the proposed model, agents in these models do not maximize a utility function [204]. In this sense, our model is more similar to the structural econometric model [23]. However, the main difference between the two is the learning process involved in the model. The proposed adoption model allows individuals to update their estimations on the characteristics of the product given the information collected from their immediate neighbors. Therefore, in our model individuals all have different estimations of the product (i.e., μ_i), whereas in [23] they all make decisions based on the same estimation.

Second, existing contagion models assume that individuals' adoption decisions are conditionally independent from that of other people in the network given the decisions of their immediate neighbors. Despite the simplicity, these models cannot capture higher-order social influence and propagation of information between neighbors, which is however captured in the proposed model via Bayesian learning. An example of such higher-order influence is shown in Figure 2-9.

Third, the critical feature that separates our study from most previous work is that an individual i continuously estimates $P_t(\boldsymbol{w}_i)$, the posterior distribution of i's perception of the product, after observing the evaluation of one of her neighbors at each time step t. This can, therefore, be considered as a dynamic information collection and updating process until a decision has been reached. Furthermore, our model captures the learning process in which individuals aggregate information from neighbors in communication network. Most existing literature, however, has overlooked

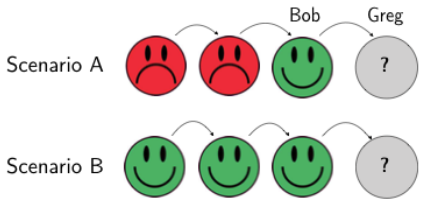
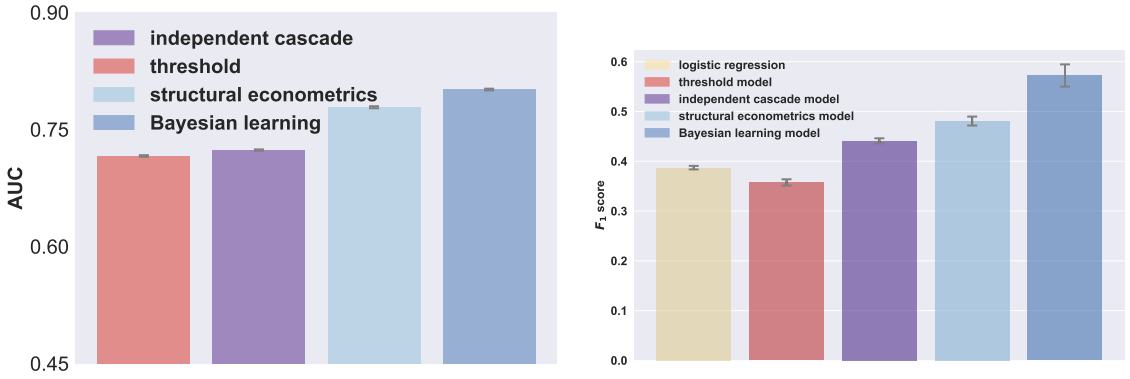


Figure 2-9: **Higher-order social influence.** Existing contagion models assume that individual behaviors are independent of the decisions of others conditioned on their immediate neighbors. That is to say, existing contagion models do not distinguish scenarios A and B in terms of the decision-making of Greg. Our model, thanks to the higher-order social influence and propagation of information between neighbors, can separate the two scenarios.

these crucial factors in people's decision-making processes. In this sense, our method is most similar to a few studies in the field of theoretical economics [3, 4, 2], in which the underlying state to be estimated can be interpreted as a single-dimension (scalar) characteristic of the product to be adopted in our case. Furthermore, these studies focus on theoretical results and do not test the validity of the models using empirical data.

2.5.3 Prediction results

To further verify the effectiveness of our proposed Bayesian learning model, we predict individual adoption behavior based on preference vectors and communication network structure by simulating the four-step process described in section A.3. We compare the performance of the proposed model with that of several baseline models: the threshold model, the independent cascade model, and the structural econometric model [23]. We evaluate the performance of the models using Area-Under-the-Curve (AUC) (i.e., the area of the curve by the true positive rate against false positive rate), and bootstrap 80% of the observation periods for 1000 times to obtain the standard error, and the 5% and 95% confidence interval. As shown in Figure 2-10, for both adoption behaviors, our model improves the AUC by about 11.7% and 20.2% comparing with the best ones in the baselines for attending the cultural performance and visiting the grocery



(a) attending cultural performance

(b) visiting retail store.

Figure 2-10: **Performance of different models in predicting adoption behavior: (a) attending cultural performance; (b) visiting retail store. The error bars correspond to the 5% and 95% confidence intervals.**

store respectively[3] compared to the best performing baseline model. This result highlights the effectiveness of modeling a local information aggregation mechanism for understanding the individual decision-making process.

2.6 Managerial implications

Firms and organizations should adapt their marketing strategies to the advancement of new technologies, such as mobile phones and social media. The present section of my thesis demonstrates that the understanding of how adoption decisions are influenced by local learning and information update process is vital to marketing firms and organizations. Our findings have the following implications on the strategies for marketing campaigns, and more broadly, for the management of large-scale behavioral change.

Firms should consider the long-range influence of individuals when measuring the importance of individuals. Our empirical results on the social influence

[3] The improvement of AUC is computed by the absolute increases in AUC normalized by the room for improvement, which can be computed as $\frac{\text{AUC}_{\text{proposed}}-\text{AUC}_{\text{benchmark}}}{1-\text{AUC}_{\text{benchmark}}}$. The structural econometric model is the best benchmark in both cases with AUC scores as 0.777 and 0.770, where our method reaches 0.803 and 0.817 respectively for attending the cultural performance and visiting the grocery store respectively.

highlight the long-range effect of individuals. This highlights the importance of considering one's higher-order social influence of individuals due to the propagation of information. Moreover, the differences in the depth of propagation and difference decay patterns of the two empirical settings also suggest that firms should perform some initial tests to estimate the effect of the social influence of different hop indices.

How should firms seed individuals given the local learning and information aggregation process? Network-based seeding is the strategy of spreading information to individuals who are "well-connected" for marketing purpose. To this end, existing node centrality measures, such as the degree, eigenvector, and Katz centralities, can be used to quantify the connectedness of individuals. However, our study demonstrates that position in the network alone is not sufficient to ensure successful marketing campaigns, as social influence can be negative due to the two mechanisms revealed by our model. It is, therefore, important to take into account the local learning process when measuring one's centrality in the network, in order to inform more effective marketing campaign designs.

How should firms design interventions to leverage the social relationship to increase adoption? Another insight from our model is that positive word-of-mouth may not necessarily increase revenue due to the potential negative influence due to the mechanisms metioned in section A.3. Whether individuals' likelihood of adoption increases or not is also influenced by the similarities between the focal individuals and their neighbors. Therefore, when firms design interventions, they need to consider both the positive and negative influences of individuals. Moreover, our results highlight the importance of finding niche network communities for synergistic local effects.

2.7 Discussion

Understanding peer influence effect on adoption behaviors has been hampered by the lack of theoretical and practical tools that can disentangle the main factors behind

the individual behavior. By considering revealed preferences from mobility patterns, the behavioral matching framework using the large-scale data enables dynamical monitoring of changes in individual taste and habit, therefore providing more robust information compared to the relatively more static socio-demographics on which classic RCT-based techniques rely. Besides, large-scale socio-demographic information is often unavailable due to privacy concerns. Thus, both robust information and broader availability are advantages of applying the proposed framework to establish causal relations in large-scale behavioral studies, which remains a difficult task for traditional techniques.

Our results reveal the subtle and often invisible effect of social influence, specifically via phone communication, on decision-making, an effect that surprisingly persists to more than three degrees of separation in the network. The persistence of the effect of social influence on decision-making, which is analogous to the physical phenomenon of ripples expanding across the water, extends the existing theories of "six degrees of separation" and "three degrees of influence" and highlights the hidden connections among behaviors of seemingly independent individuals. This long-range effect may have important implications in such domains as viral marketing, public health management, political campaigns, and social mobilization, where one may wish to leverage offline communication chains to exploit the hidden influence among individuals.

Unlike the independent cascade model, the threshold model, or the structural econometric model, we propose a Bayesian learning model to capture the underlying process of decision-making based on continuously updated posterior beliefs towards the adoption behavior. More specifically, by modeling an individual's communication and cognitive learning process dynamically, our framework introduces a local information aggregation mechanism. Interestingly, this mechanism seems to be related to the empirical decaying pattern of the effect of social influence, and may explain the superior performance of the proposed model over several baselines in predicting individual adoption behavior. This highlights the importance of incorporating individual decision-makers' dynamic local information processing in a predictive model.

One limitation of our study is that, due to the lack of ground-truth information,

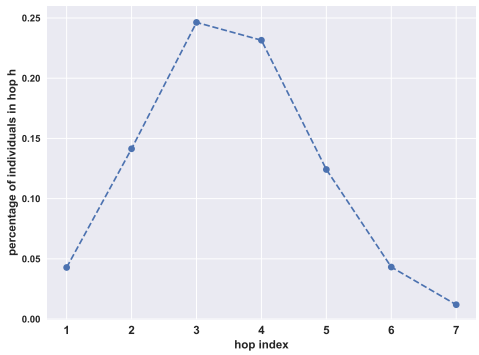
we consider individuals who were connected to the cell towers near the performance venue or store as initial adopters. This may have included people who passed by the relevant locations without actually attending the performance or visiting the store. Furthermore, since the social distance is defined as the minimum length of the shortest path between an individual and any initial adopter, we effectively consider a "strongest treatment" in estimating the treatment effect. Taking into account a multiplicative effect in case of more than one communication path (hence the possibility of multiple treatments) is thus an interesting future direction. We hope that our study can spawn a further interest in understanding costly offline adoption decisions and the contagious behavior in the social networks using other types of behavioral data.

Our work opens new possibilities in understanding social influence and contagion in offline behaviors concerning both mathematical modeling and experimental studies. The proposed Bayesian learning model can be used as a building block for further studies on, for example, developing a centrality measure in the social network that takes into account local decision-making process, or conducting counter-factual simulations to maximize social influence for behavioral change, e.g., through incentivizing key individuals to generate a local synergistic effect.

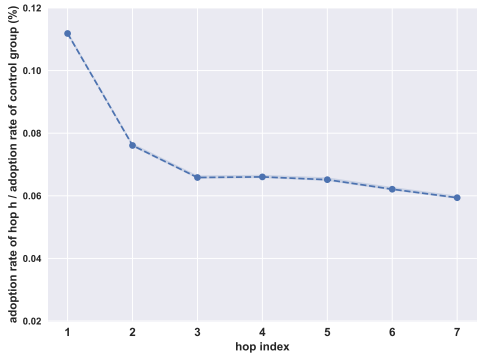
Additional analysis

Adoption rates for each hop

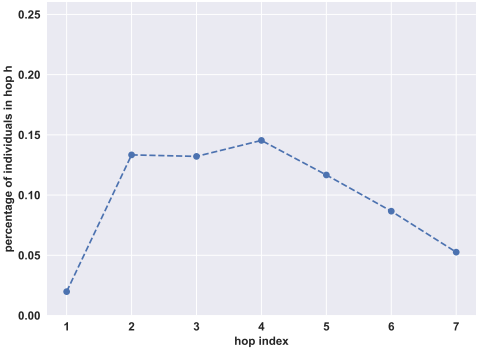
In Figure 2-11, we show the percentage of individuals as well as their adoption rates at each hop by considering the information cascades from all observation periods, for attending a cultural performance and visiting a retail store, respectively.



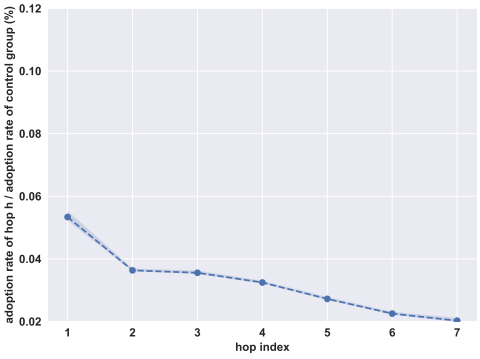
(a) attending cultural performance.



(b) attending cultural performance.



(c) visiting retail store.



(d) visiting retail store.

Figure 2-11: **Percentage of individuals (a,c) as well as their adoption rates (b,d) at each hop, computed using information cascades from all observation periods.**

Model estimation

Our model has two parameters regarding the prior perception, i.e., the priors $\boldsymbol{\mu}_{i,0}$ and $\boldsymbol{\Sigma}_{i,0}$, which we need to estimate. To reduce the degree of freedom and avoid overfitting, we set $\boldsymbol{\mu}_{i,0} = a \cdot \mathbf{1} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}_{i,0} = \text{diag}(b \cdot \mathbf{1}) \in \mathbb{R}^{n \times n}$, where a and b are the prior mean and variance of the perception of product characteristics in time instance 0. $\text{diag}(\cdot)$ is a diagonal matrix that takes the vector as its diagonal. As explained previously, we choose $n = 14$ for the adoption of attending the cultural performance and $n = 44$ for the adoption of visiting the retail store. We then choose the parameters that maximize the averaged AUC obtained by comparing the AUC computed by the ground-truth and predicted adoption decisions for all observation periods $T \in \Psi$. The motivation for choosing the AUC score instead of the accuracy is that the distribution of the cases of adoption and non-adoption is highly imbalanced.

Our model estimation proceeds with the following two steps. In the first step, we discretize the parameter space Θ and search over the entire set of possible parameters. For each possible choice of $\theta = \{a, b, \beta, p_a, p_n\} \in \Theta$, we simulate 100 times over all the information cascades in $\mathcal{C}_T$ and choose the parameter θ that maximizes the following objective function:

$$\hat{\theta} = \text{argmax} \frac{1}{|\Psi|} \sum_{T \in \Psi} \text{AUC}_T(\theta), \tag{2.10}$$

where $\text{AUC}_T(\theta)$ is the AUC obtained by comparing the ground-truth and predicted (using the parameter θ) adoption decisions for a given observation period T. The parameters are shared across all observational periods. In the second step, we estimate the distribution of $\hat{\theta}$ using a Bayesian block-bootstrap algorithm [25], which enables us to estimate the standard error of the AUC. The estimated Θ for attending the cultural performance is $\theta = \{a = 2^{13.7}, b = 2^{-3.5}, \beta = 2^1 2, p_a = 0.04, p_n = 0.015\}$, and that for visiting the retail store is $\theta = \{a = 2^6, b = 2^{-8.5}, \beta = 2^{5.4}, p_a = 0.95, p_n = 0\}$.

Chapter 3

Contextual centrality: going beyond network structure

chap:cc **Centrality is a fundamental network property that ranks nodes by their structural importance. However, the network structure alone may not predict successful diffusion in many applications, such as viral marketing and political campaigns. We propose contextual centrality, which integrates structural positions, the diffusion process, and, most importantly, relevant node characteristics. It nicely generalizes and relates to standard centrality measures. We test the effectiveness of contextual centrality in predicting the eventual outcomes in the adoption of microfinance and weather insurance. Our empirical analysis shows that the contextual centrality of first-informed individuals has higher predictive power than that of other standard centrality measures. Further simulations show that when the diffusion occurs locally, contextual centrality can identify nodes whose local neighborhoods contribute positively. When the diffusion occurs globally, contextual centrality signals whether diffusion may generate negative consequences. Contextual centrality captures more complicated dynamics on networks than traditional centrality measuresand has significant implications for network-based interventions.**[1]

[1]This work is joint with Xiaowen Dong, Junfeng Wu, and Alex Pentland.

3.1 Introduction

Individuals, institutions, and industries are increasingly connected in networks where the behavior of one individual entity may generate a global effect [109, 139, 6]. Centrality is a fundamental network property that captures an entity's ability to impact macro processes, such as information diffusion on social networks [109], cascading failures in financial institutions [6], and the spreading of market inefficiencies across industries [139]. Many interesting studies have found that the structural positions of individual nodes in a network explain a wide range of behaviors and consequences. Degree centrality predicts who is the first to be infected in a contagion [65]. Eigenvector centrality corresponds to the incentives to maximize social welfare [82]. Katz centrality is proportional to one's power in strategic interactions in network games [20]. Diffusion centrality depicts an individual's capability of spreading in information diffusion [24]. These centrality measures operate similarly, aiming to reach a large crowd via diffusion, and are solely dependent on the network structure.

However, several pieces of empirical evidence show that reaching a large crowd may decrease the evaluations of the qualities of the products. For example, sales on Groupon [55] and public announcements of popular items on Goodreads [124] are effective strategies in reaching a larger number of customers. However, both studies show that the evaluations of online reviews are negatively affected as a consequence. This phenomenon can be explained by the fact that the increasing popularity will reach individuals who hold negative opinions, and hence, translate into less favorable evaluations of quality. Let us further consider two motivating examples to demonstrate the importance of accounting for the evaluations of the nodes, and more broadly, nodal characteristics.

Example 1. Viral marketing. During a viral marketing campaign, the marketing department aims to attract more individuals to adopt the focal product. If we have ex-ante information about the customers' evaluation of the product or the likelihood of adoption, we can utilize this information to better target individuals who have higher chances of adoption and avoid wasting resources on others.

Example 2. Political campaigns. Typical Get-Out-The-Vote (GOTV) campaigns include direct mail, phone calls, and social-network advertisement [97, 44]. However, rather than simply encouraging every voter to get out the vote, a GOTV strategy should target voters who are more likely to vote for the campaigner's candidate.

In this section of my thesis, we introduce contextual centrality, which builds upon diffusion centrality proposed in Banerjee et al. and captures relevant node characteristics in the objective of the diffusion [23, 24]. Diffusion centrality focuses on the diffusion process and maximizes the number of individuals who receive the information. In other words, nodes are homogeneous. Contextual centrality is able to integrate the heterogeneity of nodes and aggregates the characteristics over one's neighborhood; hence it can be used in applications in which reaching different nodes contributes differently to the policy-makers and campaigners. In other words, it generalizes and nests degree, eigenvector, Katz, and diffusion centrality. When the spreadability (the product between the diffusion rate p and the largest eigenvalue λ_1 of the adjacency matrix) and the diffusion period T are large, contextual centrality linearly scales with eigenvector, Katz, and diffusion centrality. The sign of the scale factor is determined by the joint distribution of nodes' contributions to the objective of the diffusion and their corresponding structural positions.

We perform an empirical analysis of the diffusion of microfinance and weather insurance showing that the contextual centrality of the first-informed individuals better predicts the adoption decisions than that of the other centrality measures mentioned above. Moreover, simulations on the synthetic data show how network properties and node characteristics collectively influence the performance of different centrality measures. Further, we illustrate the effectiveness of contextual centrality over a wide range of diffusion rates with simulations on the real-world networks and relevant node characteristics in viral marketing and political campaigns.

3.2 Contextual centrality

Given a set of N individuals, the adjacency matrix of the network is $\mathbf{A}$, an N-by-N symmetric matrix. The entry A_{ij} equals 1 if there exists a link between node i and node j, and 0 otherwise. Let $\mathbf{D} = \text{diag}(\mathbf{d})$, where $d_i = \sum_{j=1}^{N} A_{ij}$ denotes the degree of node i. With Singular Value Decomposition, we have $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{\Lambda} = \text{diag}\{\mathbf{\Lambda}\} = \{\lambda_1, \lambda_2, ..., \lambda_n\}$ in a descending order and the corresponding eigenvectors are $\{\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_n\}$ with $\mathbf{U}_1$ being the leading eigenvector. We let $\circ$ denote the Hadamard product (i.e., element-wise multiplication). We use bold lowercase variables for vectors and bold upper case variables for matrices.

The diffusion process in this section of my thesis follows the independent cascade model [116]. It starts with an initial active seed. When node u becomes active, it has a single chance to activate each currently inactive neighbor v with probability P_{uv}, where $\mathbf{P} \in \mathbb{R}^{N \times N}$. We follow the terminology by Koschutzki to categorize degree, eigenvector, and Katz centrality as reachability-based centrality measures [122]. Reachability-based centrality measures aim to score a certain node v by the expected number of individuals activated if v is activated initially, $s(v, \mathbf{A}, \mathbf{P})$, and hence tend to rank higher the nodes that can reach more nodes in the network. In particular,

$$s(v, \mathbf{A}, \mathbf{P}) = \sum_{i}^{N} r_i(v, \mathbf{A}, \mathbf{P}), \tag{3.1}$$

where $r_i(v, \mathbf{A}, \mathbf{P})$ denotes the probability that i is activated if v is initially activated[116, 197, 143]. In practice, $s(v, \mathbf{A}, \mathbf{P})$ is hard to estimate. Different reachability-based centrality measures estimate it in different ways.

Diffusion centrality extends and generalizes these standard centrality measures [23]. It operates on the assumption that the activation probability of an individual i is correlated with the number of times i "hears" the information originating from the individual to be scored. Diffusion centrality measures how extensively the information spreads as a function of the initial node [23]. In other words, diffusion centrality scores node v by the expected number of times some piece of information originating from v

is heard by others within a finite number of time periods T , $s'(v, \mathbf{A}, \mathbf{P}, T)$,

$$s'(v, \mathbf{A}, \mathbf{P}, T) = \sum_{i}^{N} r'_i(v, \mathbf{A}, \mathbf{P}, T), \tag{3.2}$$

where $r'_i(v, \mathbf{A}, \mathbf{P}, T)$ is the expected number of times individual i receives the information if v is seeded. Eq. (3.2) has at least two advantages over Eq. (3.1). First, $r'_i(v, \mathbf{A}, \mathbf{P}, T)$ is computationally more efficient than tedious simulations to get $r_i(v, \mathbf{A}, \mathbf{P})$. Second, it nests degree, eigenvector, and Katz centrality [24][2].

Both Eq. (3.1) and (3.2) assume that individuals are homogeneous and contribute equally to the objectives if they have been activated. However, in many real-world scenarios, such as the two examples mentioned above, the payoff for the campaigner does not grow with the size of the cascade. Instead, different nodes contribute differently. Formally, let y_i be the contribution of individual i to the cascade payoff upon being activated. Note that y_i is context-dependent and is measured differently in different scenarios. For example, in a market campaign, y_i can be i's likelihood of adoption. In a political campaign, y_i can be the likelihood that i votes for the campaigner's political party[3]. With the independent cascade model, an individual v should be scored according to the cascade payoff if v is first-activated, $s_c(v, \mathbf{A}, \mathbf{p})$. With this, we present the following equation as a generalization and extension to

[2]It is worth noting that Eq. (3.1) and (3.2) differ in a couple of ways. First, since $r'_i(v, \mathbf{A}, \mathbf{P}, T)$ is the expected number of times i hears a piece of information, it may exceed 1. Meanwhile, since $r_i(v, \mathbf{A}, \mathbf{P})$ is the probability that i receives the information, it is bounded by 1. Second, in independent cascade model, each activated individual has a single chance to activate the non-activated neighbors. However, with the random walks of information transmission in contextual centrality, each activated individual has multiple chances with decaying probability to activate their neighbors.

[3]For the political campaign experiment in Turkey, we use individual home and work locations to build a network and regional voting data on sampling voting outcomes to use as $\mathbf{y}$. Individuals belonging to the same home neighborhood are connected according to the Watts-Strogatz model with a maximum of 10 neighbors. Same for the work neighborhoods. These two networks are superimposed to form the final network. Since we do not know the political voting preferences on an individual level, individual voting outcomes are sampled to match voting data on a regional level. Specifically, we let the actual fraction of the population that voted for the AK Party in an individual's home neighborhood be the probability that an individual votes for the AK Party. We let $y_i = +1$ represent a vote for AK party and $y_i = -1$ represent a vote for any other party. We sample a new set of voting outcomes from the regional voting distributions for each diffusion simulation.

Eq. (3.1) with heterogeneous $\mathbf{y}$,

$$\text{cascade payoff: } s_c(v, \mathbf{A}, \mathbf{p}) = \sum_i^N r_i(v, \mathbf{A}, \mathbf{P}) y_i. \tag{3.3}$$

Similar to diffusion centrality, we score nodes with the following approximated cascade payoff, $s'_c(v, \mathbf{A}, \mathbf{p}, T)$, with heterogeneous $\mathbf{y}$,

$$\text{approximated cascade payoff: } s'_c(v, \mathbf{A}, \mathbf{P}, T) = \sum_i^N r'_i(v, \mathbf{A}, \mathbf{P}, T) y_i. \tag{3.4}$$

This formulation generalizes diffusion centrality and inherits its nice properties in nesting existing reachability-based centrality measures. Moreover, it is easier to compute than Eq. (3.3)[4]. With this scoring function, we now formally propose contextual centrality.

Definition 3.2.1. *Contextual centrality (CC) approximates the cascade payoff within a given number of time periods T as a function of the initial node accounting for individuals' contribution to the cascade payoff.*

$$CC(\mathbf{A}, \mathbf{P}, T, \mathbf{y}) := \sum_{t=0}^{T} (\mathbf{P} \circ \mathbf{A})^t \mathbf{y}, \tag{3.5}$$

Heterogeneous diffusion rates across individuals are difficult to collect and estimate in real-world applications. Therefore, in the following analysis, we assume a homogeneous diffusion rate (p) across all edges. Hence, $\mathbf{P} \circ \mathbf{A}$ in Eq. (3.5) is reduced to $p\mathbf{A}$. Similar to diffusion centrality, contextual centrality is a random-walk-based centrality, where $(p\mathbf{A})^t$ measures the expected number of walks of length t between each pair of nodes and T is the maximum walk-length considered. Since T is the longest communication period, a larger T indicates a longer period for diffusion (e.g., a movie that stays in the market for a long period). In contrast, smaller T indicates a

[4]The computational complexity of the algorithm to score according to Eq. (3.3) is $O(NM^T)$, where M is the average degree, and T is the lengths of the paths that have been inspected. Note that the computational complexity of the formulation (3.5) is $O(NMT)$. We repeat the operation of multiplying a vector of length N with a sparse matrix, which has an average of M entries per row for T times. This significantly reduces the run time.

shorter diffusion period (e.g., a coupon that expires soon). On the one hand, when $p\lambda_1$ is larger than 1, CC approaches infinity as T grows. On the other hand, when $p\lambda_1 < 1$, CC is finite for $T = \infty$, which corresponds to a lack of virality, expressed in a fizzling out of the diffusion process with time. We can then specific the value of $p\lambda_1$ to bound the maximum possible CC, given the norm of the score vector $\mathbf{y}$. As presented in proposition 3.5.1 in the Supporting Information, the upper bound for CC grows with $p\lambda_1$ and the norm of the score vector.

Let us further illustrate the relationship between CC and diffusion centrality, DC for short[5]. We can represent $\mathbf{y}$ as, $\mathbf{y} = \sigma(\mathbf{y}) \cdot \mathbf{z} + \overline{\mathbf{y}} \cdot \mathbf{1}$, where $\sigma(\mathbf{y})$ and $\mathbf{z}$ are the standard deviation and the z-score normalization of $\mathbf{y}$. Using the linearity of CC with respect to $\mathbf{y}$, we can write

$$\mathrm{CC}(\mathbf{A}, p, T, \mathbf{y}) = \sigma(\mathbf{y}) \cdot \mathrm{CC}(\mathbf{A}, p, T, \mathbf{z}) + \overline{\mathbf{y}} \cdot \underbrace{\mathrm{CC}(\mathbf{A}, p, T, \mathbf{1})}_{\mathrm{DC}} \tag{3.6}$$

Eq. (3.6) shows the trade-off between the standard deviation $\sigma(\mathbf{y})$ and the mean $\overline{\mathbf{y}}$ of the contribution vector in CC. When $\overline{\mathbf{y}}$ dominates over $\sigma(\mathbf{y})$, network topology is more important in CC and it produces similar or opposite rankings to DC, depending on the sign of $\overline{\mathbf{y}}$. If the graph is Erdos-Renyi and T is small enough, then, on expectation, the term $\overline{\mathbf{y}} \cdot \mathrm{DC}$ dominates as the size of the network approaches infinity, as presented in Theorem 3.5.1 in the Supporting Information. However, when $\sigma(\mathbf{y})$ dominates over $\overline{\mathbf{y}}$, CC and DC generate very different rankings.

The relevant node characteristics ($\mathbf{y}$) provides the ex-ante estimation about one's contribution. Whether to incorporate $\mathbf{y}$ is the main difference between our centrality and existing centrality measures. In the real-world data, the observation or estimation on $\mathbf{y}$ can be noisy, biased, or stochastic. Therefore, we discuss the robustness of contextual centrality in response to perturbations in $\mathbf{y}$ in the Supporting Information.

We define the following terms, which we use throughout the project:

[5]In Banerjee et al.[23], $\mathrm{DC} = \sum_{t=1}^{T} (p\mathbf{A})^t$. To derive the following relationship between CC and DC, we add the score of reaching the first seeded individual into computing diffusion centrality. Hence, $\mathrm{DC} = \sum_{t=0}^{T} (p\mathbf{A})^t$. Adding the first seeded individual into the scoring function produces the same ranking as the one used in Banerjee et al.

- Spreadability ($p\lambda_1$) captures the capability of the campaign to diffuse on the network depending on the diffusion probability (p) via a certain communication channel, and the largest eigenvalue (λ_1) of the network.

- Standardized average contribution ($\frac{\overline{\mathbf{y}}}{\sigma(\mathbf{y})}$) is computed as the average of the contributions normalized by the standard deviation of the contribution. The sign of $\frac{\overline{\mathbf{y}}}{\sigma(\mathbf{y})}$ indicates whether the average contribution is positive or not. Moreover, the larger the magnitude of $\frac{\overline{\mathbf{y}}}{\sigma(\mathbf{y})}$, the more homogeneous the contributions are.

- Primary contribution ($\mathbf{U}_1^T\mathbf{y}$) measures the joint distribution of the structural importance and nodal contributions. It captures whether people who dominate important positions have positive contributions or not.

Properties of contextual centrality when $p\lambda_1 > 1$ and T is large.

Let us first provide the approximation of contextual centrality in this condition, which reveals one of the prominent advantages of contextual centrality. By the Perron-Frobenius Theorem, we have $|\lambda_j| \leq \lambda_1$ for every j. Moreover, if we assume that the graph is non-periodic, then in fact $|\lambda_j| < \lambda_1$ for all $j \neq 1$. Note that the typical random graph is not periodic, so this assumption is reasonable. Thus, when $p\lambda_1 > 1$, the term $(p\lambda_1)^t$ grows exponentially faster than $(p\lambda_j)^t$ for $j \neq 1$ so that the $j = 1$ term dominates for sufficiently large values of T, and we obtain the approximation for contextual centrality ($\mathrm{CC}_{\mathrm{approx}}$):

$$\mathrm{CC} = \sum_{j=1}^{n}\sum_{t=0}^{T}(p\lambda_j)^t\mathbf{U}_j\mathbf{U}_j^T\mathbf{y} \approx \mathrm{CC}_{\mathrm{approx}} = \Big(\sum_{t=0}^{T}(p\lambda_1)^t\mathbf{U}_1^T\mathbf{y}\Big)\mathbf{U}_1. \tag{3.7}$$

This approximation reveals some desirable properties of contextual centrality. Crucially, $\mathrm{CC}_{\mathrm{approx}}$ is simply a scalar multiple of the leading eigenvector when $p\lambda_1 > 1$ and T is large. Therefore, the sign of $\mathbf{U}_1^T\mathbf{y}$ determines the direction of the relationship between $\mathrm{CC}_{\mathrm{approx}}$ and eigenvector centrality. By Perron-Frobenius Theorem, all elements in this leading eigenvector are nonnegative. Thus, the approximated cascade payoff, Eq. (3.4), for seeding any individual is nonpositive if $\mathbf{U}_1^T\mathbf{y} < 0$, $p\lambda_1 > 1$ and

T is large. This shows that in this condition, the approximated cascade payoff is nonpositive for seeding any individual, so the campaigner should select a diffusion channel with a lower diffusion rate to take advantage of the local neighborhood with positive contributions. Eq. (3.7) naturally suggests the following relationships between $\mathrm{CC}_{\mathrm{approx}}$ and eigenvector centrality.

- If $\mathbf{U}_1^T\mathbf{y} > 0$, $\mathrm{CC}_{\mathrm{approx}}$ and eigenvector centrality produce the same rankings.
- If $\mathbf{U}_1^T\mathbf{y} < 0$, $\mathrm{CC}_{\mathrm{approx}}$ and eigenvector centrality produce the opposite rankings.

The approximation does not hold when $\mathbf{U}_1^T\mathbf{y} = 0$, which is also unlikely to happen in practice. Hence, we disregard this case. Similarly, we relate contextual centrality to diffusion centrality ($\mathrm{C}_{\mathrm{Diffusion}}$) and Katz centrality ($\mathrm{C}_{\mathrm{Katz}}$),

$$\begin{aligned}
\mathrm{C}_{\mathrm{Diffusion}} &\propto \sum_{t=1}^{\infty}(p\lambda_1)^t\mathbf{U}_1(\mathbf{U}_1^T\mathbf{1}) = \frac{\sum_{t=1}^{\infty}(p\lambda_1)^t(\mathbf{U}_1^T\mathbf{1})}{\sum_{t=0}^{T}(p\lambda_1)^t\mathbf{U}_1^T\mathbf{y}}\mathrm{CC}_{\mathrm{approx}},\\
\mathrm{C}_{\mathrm{Katz}} &\propto \sum_{t=0}^{\infty}(\alpha\lambda_1)^t\mathbf{U}_1(\mathbf{U}_1^T\mathbf{1}) = \frac{\sum_{t=0}^{\infty}(\alpha\lambda_1)^t(\mathbf{U}_1^T\mathbf{1})}{\sum_{t=0}^{T}(p\lambda_1)^t\mathbf{U}_1^T\mathbf{y}}\mathrm{CC}_{\mathrm{approx}},
\end{aligned} \tag{3.8}$$

where α is the decay factor in Katz centrality. Similar to Eq. (3.7), all terms on the right-hand-side of Eq. (3.8) are positive except for $\mathbf{U}_1^T\mathbf{y}$, which similarly determines the direction of the relationship.

3.3 Results

3.3.1 Methods

In this study, we compare contextual centrality with diffusion centrality and other widely-adopted reachability-based centrality measures - degree, eigenvector, and Katz centrality. We compute degree centrality by taking the degree of each node, normalized by $N-1$. We compute eigenvector centrality by taking the leading eigenvector $\mathbf{U}_1$ with unit length and nonnegative entries. We compute Katz centrality as $\sum_{t=0}^{\infty}(\alpha\mathbf{A})^t\mathbf{1}$, setting α, which should be strictly less than λ_1^{-1}, to $0.9\cdot\lambda_1^{-1}$. We compute diffusion centrality as $\sum_{t=1}^{T}(p\mathbf{A})^t\mathbf{1}$. For both diffusion and contextual centrality, we set $T = 16$,

except for the microfinance in Indian villages setting, where we set T as done by Banerjee et al. [23].

Simulations of the diffusion process in each setting follow the independent cascade model [116]. For each centrality, the highest-ranked node is set to be the initial seed. We compute cascade payoff by summing up the individual contributions of all the nodes reached in the cascade. For each parameter tested in different settings, we run 100 simulations.

In the empirical analysis of microfinance in Indian villages and weather insurance in Chinese villages, we build models to predict the adoption likelihood to use as $\mathbf{y}$ in computing contextual centrality. For each setting, we use the data provided in Banerjee et al. [23] and Cai et al. [57], respectively, as inputs to a feed-forward neural network trained to predict the adoption likelihood based on the adoption decisions of first-informed individuals. For the microfinance in Indian villages, the covariates include village size, quality of access to electricity, quality of latrines, number of beds, number of rooms, the number of beds per capita, and the number of rooms per capita. For the weather insurance in Chinese villages setting, 39 of the provided characteristics are selected as inputs by choosing those for which all households had data after removing households with many missing characteristics.

For the political campaign experiment in Turkey, we use individual home and work locations to build a network and regional voting data on sampling voting outcomes to use as $\mathbf{y}$. Individuals belonging to the same home neighborhood are connected according to the Watts-Strogatz model with a maximum of 10 neighbors. Same for the work neighborhoods. These two networks are superimposed to form the final network. Since we do not know the political voting preferences on an individual level, individual voting outcomes are sampled to match voting data on a regional level. Specifically, we let the actual fraction of the population that voted for the AK Party in an individual's home neighborhood be the probability that individual votes for the AK Party. We let $y_i = +1$ represent a vote for AK party and $y_i = -1$ represent a vote for any other party. We sample a new set of voting outcomes from the regional voting distributions for each diffusion simulation.

For the synthetic setting, we generate a new random graph for each simulation, according to Barabasi-Albert, Erdos-Renyi, and Watts-Strogatz models. The size n of each graph varies between 30 and 300. For Barabasi-Albert, m varied between 1 and n. For Erdos-Renyi, p varies between 0 and 1. For Watts-Strogatz, k varies between $\ln n$ and n, and p varies between 0 and 1. Individual contributions $\mathbf{y}$ are sampled from a normal distribution with unit standard deviation. Note that the scale of $\mathbf{y}$ does not change the rankings of contextual centrality.

Predictive power of contextual centrality

We study two real-world empirical settings, adopting microfinance in 43 Indian villages[6] and adopting weather insurance in 47 Chinese villages[7]. In each setting, there is a set of first-informed households in each village who went on to spread the information. We evaluate the adoption outcome of all other households in the village, which are not first-informed. We use the adoption likelihood for the contribution vector $\mathbf{y}$ in computing contextual centrality, which is predicted using a model based on the adoption decisions of the first-informed households[8]. Similar to Banerjee et al. [23], we evaluate the R^2 of a linear regression model for both settings. The independent variables include the average centrality of first-informed households and the village size, a control variable. The dependent variable is the fraction of non-first-informed households in a village which adopted.

In Fig. 3-1, we show how the R^2 for various measures of centrality varies with $p\lambda_1$, in which the choice of p influences the two centrality measures that account

[6]The data is made public by Banerjee et al. [23].

[7]The data is made public by Cai et al. [57].

[8]In the empirical analysis of both settings, we build models to predict the adoption likelihood for each individual to use as $\mathbf{y}$ in computing contextual centrality. For each setting, we use the data provided in Banerjee et al. [23] and Cai et al. [57], respectively, as inputs to a feed-forward neural network trained to predict the adoption likelihood based on the adoption decisions of first-informed individuals. Hyperparameters, including hidden layers, activation function, and regularization, were tuned using grid search with 10-fold cross-validation. For the microfinance in Indian villages, the covariates include village size, quality of access to electricity, quality of latrines, number of beds, number of rooms, the number of beds per capita, and the number of rooms per capita. For the weather insurance in Chinese villages setting, 39 of the provided characteristics are selected as inputs by choosing those for which all households had data after removing households with many missing characteristics.

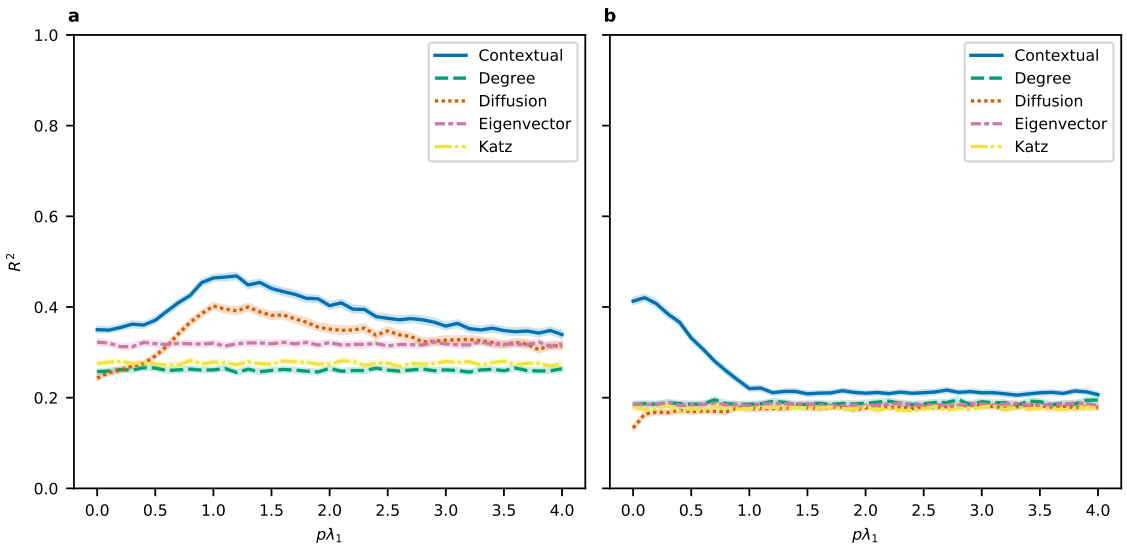


Figure 3-1: **Predictive power of contextual centrality. We show how the average centrality of first-informed individuals predicts the eventual adoption rate of non-first-informed individuals in (a) microfinance and (b) weather insurance.** The y-axis shows the 95% confidence interval of R^2 computed from 1000 bootstrap samples from ordinary least squares regressions controlling for village size. The x-axis shows varying values for $p\lambda_1$, which influences only diffusion centrality and contextual centrality.

for the diffusion process - diffusion centrality and contextual centrality. We see that the contextual centrality outperforms all other standard centrality measures[9], which indicates that marketing campaigners or social planners will benefit from using contextual centrality as the seeding strategy to maximize participation. This result also highlights that utilizing ex-ante information about customers' likelihood of adoption helps to design better targeting strategies. Similar results without control variables and with more control variables are presented in the Supporting Information as a robustness check.

[9]In this study, we compare contextual centrality with diffusion centrality and other widely-adopted reachability-based centrality measures – degree, eigenvector, and Katz centrality. We compute degree centrality by taking the degree of each node, normalized by $N-1$. We compute eigenvector centrality by taking the leading eigenvector $\mathbf{U}_1$ with unit length and nonnegative entries. We compute Katz centrality as $\sum_{t=0}^{\infty}(\alpha\mathbf{A})^t\mathbf{1}$, setting α, which should be strictly less than λ_1^{-1}, to $0.9\cdot\lambda_1^{-1}$. We compute diffusion centrality as $\sum_{t=1}^{T}(p\mathbf{A})^t\mathbf{1}$. For both diffusion and contextual centrality, we set $T=16$, except for the microfinance in Indian villages setting, where we set T the same as Banerjee et al. [23].

Performance of contextual centrality relative to other centrality measures on random networks To better understand CC's performance with respect to different parameters ($p\lambda_1$, $\frac{\overline{\mathbf{y}}}{\sigma(\mathbf{y})}$), we next perform simulations on randomly generated synthetic networks and contribution vectors ($\mathbf{y}$). For the synthetic setting, we generate a new random graph for each simulation, according to Barabasi-Albert, Erdos-Renyi, and Watts-Strogatz models. The size of n of each graph varies between 30 and 300. For Barabasi-Albert, m varied between 1 and n. For Erdos-Renyi, p varies between 0 and 1. For Watts-Strogatz, k varies between $\ln n$ and n, and p varies between 0 and 1. Individual contributions $\mathbf{y}$ are sampled from a normal distribution with unit standard deviation. Note that the scale of $\mathbf{y}$ does not change the rankings of contextual centrality. Simulations of the diffusion process in each setting follow the independent cascade model [116]. For each centrality, the highest-ranked node is set to be the initial seed. We compute cascade payoff by summing up the individual contributions of all the nodes reached in the cascade. For each parameter tested in different settings, we run 100 simulations.. To compare the performance of contextual centrality against the other centrality measures, we use "relative change" (calculated as $\frac{a-b}{max(|a|,|b|)}$, where a is a given centrality's average payoff and b is the maximum average payoff of the other centrality measures)[10].

Fig. 3-2 displays the relative change between CC's average payoff and the maximum average payoff of the other centrality measures aggregated over 100 runs of simulations for varying values of $\frac{\overline{\mathbf{y}}}{\sigma(\mathbf{y})}$ and $p\lambda_1$ on three different types of simulated graphs. We can see that CC performs well when $\overline{\mathbf{y}} < 0$, $p\lambda_1 < 1$, and $\frac{\overline{\mathbf{y}}}{\sigma(\mathbf{y})}$ is small in magnitude. We will now discuss each of these cases in more detail.

When $\overline{\mathbf{y}} < 0$, maximizing the reach of the cascade is not ideal because that will result in a cascade payoff, which more closely reflects $\overline{\mathbf{y}}$. CC differs from the other centrality measures in that it does not try to maximize the reach of the cascade. Note the dark blue diagonal band present in all plots in Fig. 3-2. Since the magnitude of

[10]We chose "relative change" for comparison since it gives a sense of when the payoffs are different from the optimal centrality while keeping the magnitudes of the payoffs in perspective. This measure has some desirable properties. First, its value is necessarily between -2 and 2, so our scale for comparison is consistent across scenarios. Second, its magnitude does not exceed one unless a and b differ in sign, so we can tell if a centrality gets a positive average payoff while the rest do not

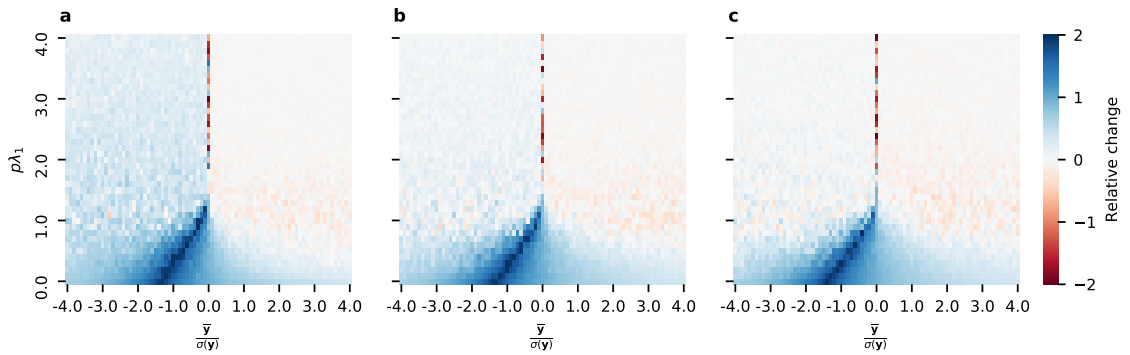


Figure 3-2: **Performance of contextual centrality relative to other centrality measures on random networks.** Each plot shows the relative change, computed as $\frac{a-b}{\max(|a|,|b|)}$ where a is CC's average payoff and b is the maximum average payoff of the other centrality measures, for varying values of $\frac{\overline{y}}{\sigma(y)}$ and $p\lambda_1$. The plots correspond to the results on random networks generated according to the (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

the relative change exceeds one only when the values being compared have opposite signs, this region shows that there are many settings where the standardized average contribution is negative. Nonetheless, CC achieves a positive average payoff while the other centrality measures do not.

When $p\lambda_1$ is small, it is essential to seed an individual whose local neighborhood has higher individual contributions since there is not much risk of diffusing to individuals with lower individual contributions[11]. This highlights CC's advantage in discriminating the local neighborhoods with positive payoffs from those with negative payoffs while the other centrality measures cannot.

When $\frac{\overline{y}}{\sigma(y)}$ is small in magnitude, CC takes advantage of the greater relative variations between contributions. As $\frac{\overline{y}}{\sigma(y)} \to +\infty$, Eq. (3.6) tells us that CC will seed similar to DC, which explains why CC loses some of its advantage. However, as $\frac{\overline{y}}{\sigma(y)} \to -\infty$, Eq. (3.6) tells us that CC will seed opposite to DC, which explains why CC maintains an advantage.

We now discuss the regions where CC does not seem to offer an advantage. Note that parameters for which CC's average payoff is lower than that of some other

[11] As an extreme case, consider $p\lambda_1 = 0$. In this case, the diffusion rate is 0, so seeding an individual with a high individual payoff makes much more sense than seeding an individual with high topological importance.

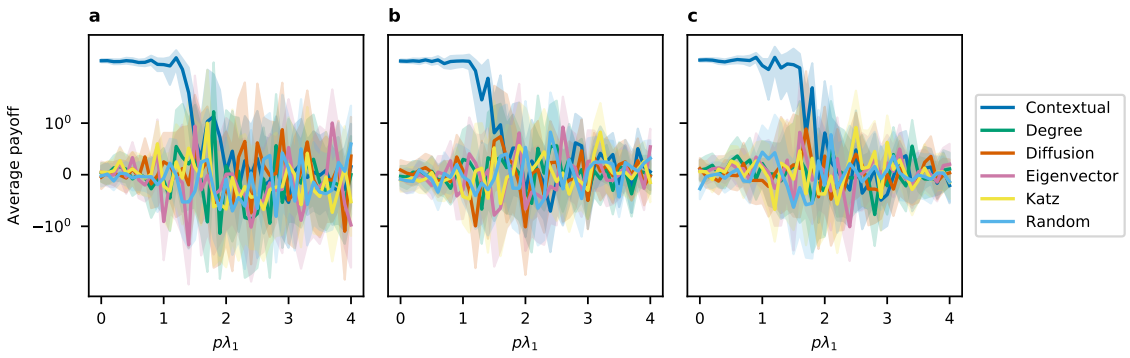


Figure 3-3: **Average payoffs when standardized average contribution is 0.** Here we show the average payoff with 95% confidence interval when seeding with different methods on (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

centrality often neighbor similar parameters for which CC's average payoff is the same, or sometimes higher, than those of the other centrality measures. This suggests that CC is performing comparably, which is what we expect as $p\lambda_1$ increases since the initial seed matters less as the diffusion process reaches more individuals. In Fig. 3-3 and Fig. 3-4, we show the average payoffs of different seeding methods with 95% confidence interval when the standardized average contribution is 0 and 1, respectively, on (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models. Note that when $p\lambda_1$ is small, CC dominates the other seeding methods. As $p\lambda_1$ increases, CC's performance is on par with other centrality measures, as can be seen from the highly overlapping confidence intervals. This pattern holds for other values of the standardized average contribution. Similar figures to Fig. 3-3 and Fig. 3-4 for other values of the standardized average contribution can be found in the Supporting Information.

Performance of contextual centrality relative to other centrality measures on real-world networks Next, we analyze the performance of contextual centrality in achieving the cascade payoff, as defined in Eq. (3.3), using simulations on three real-world settings, namely adoption of microfinance, adoption of the weather insurance, and political voting campaign, as shown in Fig. 3-5. To compare the performance of contextual centrality against the maximum of centrality measures for each condition,

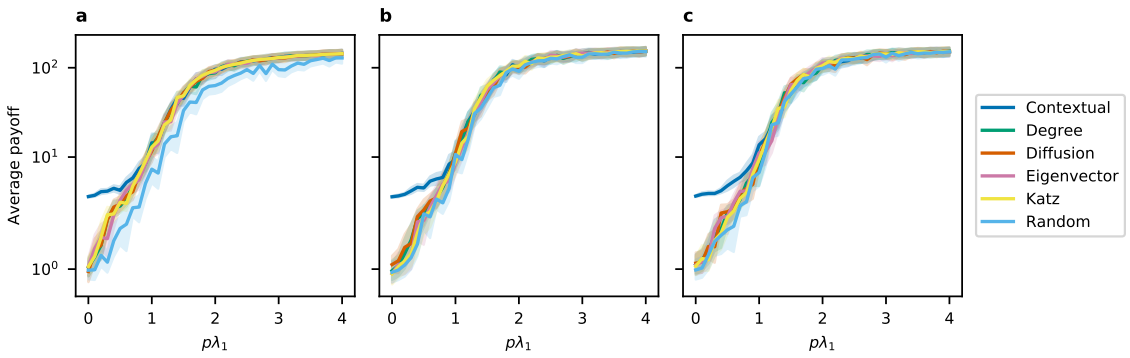


Figure 3-4: **Average payoffs when standardized average contribution is 1.** Here we show the average payoff with 95% confidence interval when seeding with different methods on (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

we use "relative change" as before. We observe the network structure ($\mathbf{A}$) and adoption decisions in the campaign for microfinance and weather insurance. In the campaign for political votes, we generate the network structure and the contribution vector from the empirical distributions. We vary the diffusion rate of p in the independent cascade model to examine how it influences the performances of different centrality measures. We see that in (a) campaign for microfinance and (b) campaign for weather insurance, CC outperforms the other centrality measures when $p\lambda_1$ is small. While in (c) campaign for political votes, CC outperforms the other centrality measures for all $p\lambda_1$. The standardized average contributions of (a), (b), and (c) are 2.29, 5.27, and -2.22, respectively. This result is consistent with the results presented in Fig. 3-2. It shows that contextual centrality can greatly outperform other centrality measures when the standardized average contribution is negative for a wide range of $p\lambda_1$. When standardized average contribution is positive, contextual centrality outperforms other centrality measures when the spreadability is small and achieves comparable results with other centrality measures as the spreadability further increases.

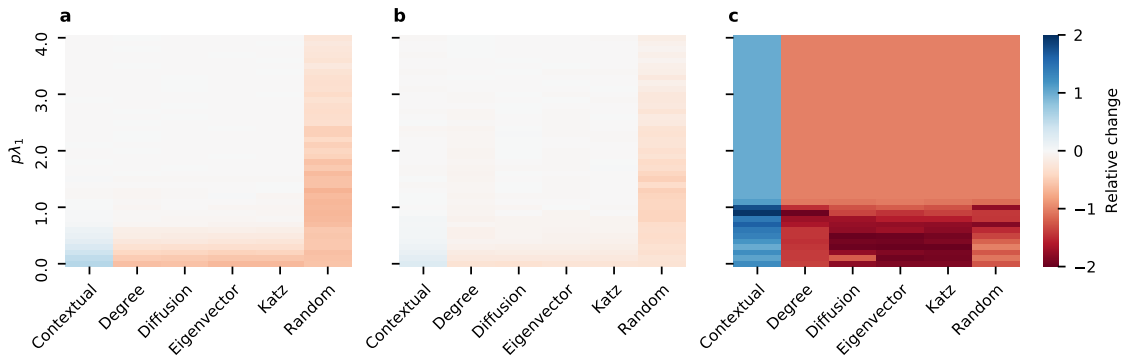


Figure 3-5: **Performance of contextual centrality relative to other centrality measures on real-world networks, including (a) microfinance, (b) weather insurance, and (c) political campaign.** Each plot shows the relative change for varying values of $p\lambda_1$. We compare contextual centrality with degree centrality, diffusion centrality, eigenvector centrality, Katz centrality, and random seeding.

Approximation of contextual centrality and the importance of primary contribution

A negative contextual centrality score indicates that seeding with the particular node will generate a negative payoff. Therefore, we design a seeding strategy in which we seed only if the maximum of contextual centrality is nonnegative. As shown by the blue dashed and solid lines in Fig. 3-6, the new seeding strategy, "seed nonnegative", performs better than always seeding the top-ranked individual. Building upon Eq. (3.7), we introduce a variation of eigenvector centrality, "eigenvector adjusted", as the product of eigenvector centrality and the primary contribution ($\mathbf{U}_1^T\mathbf{y}$). This variation of eigenvector centrality performs on par with contextual centrality as $p\lambda_1$ grows large as expected according to Eq. (3.7). "Eigenvector adjusted" greatly outperforms eigenvector centrality[12]. Comparing the strategies in Fig. 3-6, the new strategy of accounting for the sign of the centrality measures improves the average payoffs by an order of magnitude. This pattern also highlights the importance of the

[12]Another variation of eigenvector centrality is to adjust eigenvector centrality by $\overline{\mathbf{y}}$. Note that the sign of $\mathbf{U}_1^T\mathbf{y}$ does not always equal $\overline{\mathbf{y}}$. When the signs differ, seeding only when $\mathbf{U}_1^T\mathbf{y}$ is positive produces a higher cascade payoff when $p\lambda_1$ is not too large. However, as $p\lambda_1$ further increases and the diffusion saturates most of the network, the sign of $\overline{\mathbf{y}}$ predicts that of the cascade payoff. However, larger $p\lambda_1$ is not as interesting as smaller ones, which happens more frequently in real life. We present average cascade payoff comparing the two strategies when $\overline{\mathbf{y}}(\mathbf{U}_1^T\mathbf{y}) < 0$ in the section 3.6.

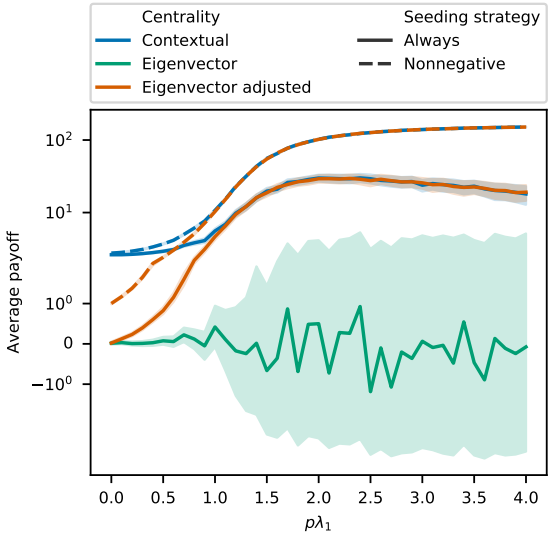


Figure 3-6: **Average cascade payoff for variations of contextual centrality and eigenvector centrality. The x-axis is $p\lambda_1$, and the y-axis is the average payoff, with the shaded region as the 95% confidence interval.** For "eigenvector adjusted" centrality, we multiply eigenvector centrality with the primary contribution $\mathbf{U}_1^T\mathbf{y}$. For "seed nonnegative", we only seed if the maximum of the centrality measure is nonnegative, otherwise it is named "seed always".

primary contribution to campaign strategies. We present figures for the analogous variations of the other centrality measures in the Supporting Information.

Homophily and the maximum of contextual centrality

Homophily is a long-standing phenomenon in social networks that describes the tendency of individuals with similar characteristics to associate with one another [147]. The strength of homophily is measured by the difference in the contributions of the neighbors, $\sum_{i,j}^{N} A_{ij}(y_i - y_j)^2$. We analyze the relationship between the strength of homophily and the approximated cascade payoff by seeding the highest-ranked node in contextual centrality in Fig. 3-7. After controlling for $\frac{\overline{\mathbf{y}}}{\sigma(\mathbf{y})}$ and $p\lambda_1$, we regress the maximum of the contextual centrality on the strength of homophily of the network separately for three conditions of $\frac{\overline{\mathbf{y}}}{\sigma(\mathbf{y})}$. When the spreadability of contextual centrality is small, stronger homophily tends to correlate with a large approximated cascade payoff across all graph types. This result shows that stronger homophily of the

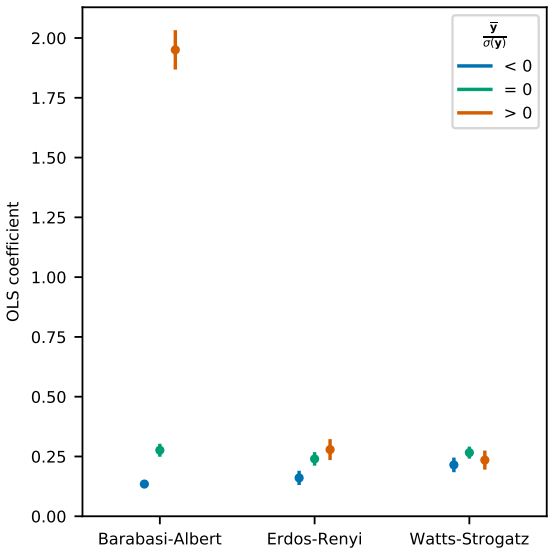


Figure 3-7: **Homophily and maximum of contextual centrality when $p\lambda_1 < 1$.** We regress the maximum of contextual centrality on homophily after controlling for $\frac{\bar{\mathbf{y}}}{\sigma(\mathbf{y})}$ and $p\lambda_1$. The y-axis is the OLS coefficients of homophily (with the vertical line as the 95% confidence interval) and the x-axis corresponds to three types of networks. We perform the analysis separately for $\frac{\bar{\mathbf{y}}}{\sigma(\mathbf{y})}$ being larger than, smaller than and equals to zero.

network predicts higher approximated cascade payoff with small spreadability. When the network is Barabasi-Albert and $\frac{\bar{\mathbf{y}}}{\sigma(\mathbf{y})} > 0$, the relationship is the strongest. As the spreadability further increases, the correlation between contextual centrality and homophily drops dramatically, and thereby we exclude the scenarios when $p\lambda_1 > 1$.

3.4 Discussion

Contextual centrality sheds light on the understanding of node importance in networks by emphasizing node characteristics relevant to the objective of the diffusion other than the structural topology, which is vital for a wide range of applications, such as marketing or political campaigns on social networks. Notably, nodal contributions to the objective, the diffusion probability, and network topology jointly produce an effective campaign strategy. It should now be clear with the thorough simulations and empirical analysis in this study that exposing a large portion of the population in the diffusion is not always desirable.

- When the spreadability is small, contextual centrality effectively ranks the nodes whose local neighborhoods generate larger cascade payoffs the highest.

- When the spreadability is large, the primary contribution tends to predict the sign of the approximated cascade payoff.

Meanwhile, for a given contribution vector ($\mathbf{y}$), the policy-maker can influence the diffusion rate to take advantage of local diffusion and locate nodes whose local neighborhood generates large cascade payoff. In practice, the policy-maker can first estimate the contribution vector ($\mathbf{y}$), and then calculate the maximum of contextual centrality for a range of $p\lambda_1$, which approximates the cascade payoff. Finally, the policy-maker can compute the optimal corresponding p given the leading eigenvector (λ_1).

When the primary contribution is negative, the campaigner might need to reduce the spreadability of the campaign to take advantage of the individuals whose local neighborhoods generate positive approximated cascade payoff in aggregation. To reduce the spreadability of the campaign, the campaigner can resort to campaign channels with lower diffusion probability and less viral features, such as direct mail.

As the standardized average contribution increases, the contribution vector becomes comparatively more homogeneous and comparatively less important than the network structure. Therefore, when the average contribution is positive, seeding with contextual centrality becomes similar to seeding with diffusion centrality.

Moreover, contextual centrality emphasizes the importance of incorporating node characteristics that are exogenous to the network structure and the dynamic process. More broadly, contextual centrality provides a generic framework for future studies to analyze the joint effect of network structure, nodal characteristics, and the dynamic process. Other than applications on social networks, contextual centrality can be applied to analyzing a wide range of networks, such as the biology networks (e.g., rank the importance of genes by using the size of their evolutionary family as the contribution vector[145]), the financial networks (e.g., rank the role of institutions in risk propagation in financial crisis with their likelihoods of failure as the contribution

vector[6]), and the transportation networks (e.g., rank the importance of airports with the passengers flown per year as the contribution vector[99]).

3.5 Properties of contextual centrality

3.5.1 Bounds and distribution of contextual centrality in terms of spreadability

In this section, we present the upper bound for maximum possible contextual centrality. When $p\lambda_1$ is larger than 1, CC approaches infinity as T grows. On the other hand, when $p\lambda_1 < 1$, CC is finite for $T = \infty$, which can be understood as a lack of virality, expressed in a fizzling out of the diffusion process with time. We can use the value of $p\lambda_1$ to bound the maximum possible CC given the norm of the score vector $\mathbf{y}$.

Proposition 3.5.1.

$$max(CC(\mathbf{A}, p, T, \mathbf{y})) \leq ||CC(\mathbf{A}, p, T, \mathbf{y})||$$

$$\leq \frac{1 - (p\lambda_1)^{T+1}}{1 - p\lambda_1}||\mathbf{y}||$$

If, in addition, $p\lambda_1 < 1$, then this is further bounded by $\frac{1}{1-p\lambda_1}||\mathbf{y}||$.

Proof. The first inequality, $max(CC(\mathbf{A}, p, T, \mathbf{y})) \leq ||CC(\mathbf{A}, p, T, \mathbf{y})||$, is clear.

Next we use the matrix norm $||\mathbf{A}|| := sup\{||\mathbf{A}x||/||x|| : x \neq 0\}$, which by definition satisfies $||\mathbf{A}x|| \leq ||\mathbf{A}|| \cdot ||x||$ for all x, and which coincides with spectral radius $\rho(\mathbf{A})$ for symmetric matrices. Since, for us, $\rho(\mathbf{A}) = \lambda_1$, we have

$$||CC(\mathbf{A}, p, T, \mathbf{y})|| = ||(\sum_{t=0}^{T}(p\mathbf{A})^t)\mathbf{y}|| \leq ||(\sum_{t=0}^{T}(p\mathbf{A})^t)|| \cdot ||\mathbf{y}||$$

$$\leq \sum_{t=0}^{T}||(p\mathbf{A})^t|| \cdot ||\mathbf{y}|| = \sum_{t=0}^{T}(p\lambda_1)^t \cdot ||\mathbf{y}|| \leq \frac{1 - (p\lambda_1)^{T+1}}{1 - p\lambda_1}||\mathbf{y}||$$

which, if $p\lambda_1 < 1$, can be further bounded by $\frac{1}{1-p\lambda_1}||\mathbf{y}||$ □

While the above result bounds contextual centrality from above, the actual value of CC is highly variable, depending on the structure of the graph and the distribution of the score vector among its nodes. For a discussion of expected CC among random networks, see the Erdos-Reyni section below. Next, we discuss the behavior of contextual centrality when $\mathbf{y}$ is variable.

3.5.2 Robustness of contextual centrality in response to perturbations in y

As discussed in the main body of this chapter, in real-world data, node characteristics can be noisy, stochastic, and biased. Therefore, it is essential to analyze the robustness of contextual centrality in response to small perturbations in $\mathbf{y}$. We first perform a sensitivity analysis, studying bounds on the error in contextual centrality in terms of noise in $\mathbf{y}$, and then study contextual centrality as a random variable assuming a multivariate normal model of $\mathbf{y}$.

Sensitivity Analysis

We let the observed (or estimated) score vector be $\hat{\mathbf{y}}$ and let $\mathbf{y}$ be the true score vector. The errors in the score vector are given by the vector $\Delta\mathbf{y} := \mathbf{y} - \hat{\mathbf{y}}$ and similarly $\Delta\text{CC} := \text{CC}(\mathbf{A}, p, T, \hat{\mathbf{y}}) - \text{CC}(\mathbf{A}, p, T, \mathbf{y})$ is the error between the CC computed from observed and actual data.

We have the following bound on $||\Delta\text{CC}||$, which follows directly from Proposition 3.5.1 and the fact that CC is linear with respect to the score vector $\mathbf{y}$.

Corollary 3.5.1.

$$||\Delta CC|| \leq \frac{1 - (p\lambda_1)^{T+1}}{1 - p\lambda_1}||\Delta\mathbf{y}||$$

If, in addition, $p\lambda_1 < 1$*, then this is further bounded by* $\frac{1}{1-p\lambda_1}||\Delta\mathbf{y}||$.

This shows that when $p\lambda_1 < 1$, then as long as the error in $\mathbf{y}$ is sufficiently small, the error in CC will be small as well. However, the larger $p\lambda_1$ is, the more a small error in $\mathbf{y}$ can become amplified as an error in CC.

Next we focus on the case that $p\lambda_1 > 1$. In this case, we have shown in the main body of this chapter that for large T, contextual centrality is well-approximated by $(\mathbf{U}_1^T\mathbf{y})\mathbf{U}_1$, where $\mathbf{U}_1$ is the eigenvector with the largest eigenvalue. Thus, in this case, the primary contribution $\mathbf{U}_1^T\mathbf{y}$ is an essential quantity whose sign roughly determines the relative ranking of contextual centrality. Hence, we analyze its sensitivity to noise in $\mathbf{y}$. The error in primary contribution is simply $\mathbf{U}_1^T\Delta\mathbf{y}$, whose magnitude is bounded by $||\Delta\mathbf{y}||$. Thus if $\Delta\mathbf{y}$ is small enough so that $||\Delta\mathbf{y}|| < \mathbf{U}_1^T\hat{\mathbf{y}}$, this perturbation will not affect the sign of the primary contribution, so the relative ranking in CC will tend to stay fixed. Otherwise, the relative ranking is at risk of flipping.

Contextual centrality as a random variable

Next, to study the impact of stochasticity in $\mathbf{y}$, we suppose that $\mathbf{y}$ is a multivariate random variable with mean vector $\hat{\mathbf{y}}$ and covariance matrix Σ. Let $\mathbf{B} := \sum_{t=0}^{T}(p\mathbf{A})^t$. Since $\mathrm{CC}(\mathbf{A}, p, T, \mathbf{y}) = \mathbf{B}\cdot\mathbf{y}$ is a linear transformation of the multivariate normal variable $\mathbf{y}$, it is also a multivariate normal variable, with mean $\mathbf{B}\hat{\mathbf{y}} = \mathrm{CC}(\mathbf{A}, p, T, \hat{\mathbf{y}})$ and covariance matrix $\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}$.

To simplify, consider the case that $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, that is, the y_i are uncorrelated and have the same standard deviation σ. Then the covariance matrix of $\mathrm{CC}(\mathbf{A}, p, T, \mathbf{y})$ is $\sigma^2\mathbf{B}^2$.

That is, we have

$$\mathrm{Cov}(\mathrm{CC}(\mathbf{A}, p, T, \mathbf{y})_i, \mathrm{CC}(\mathbf{A}, p, T, \mathbf{y})_j) = \sigma^2(\mathbf{B}e_i)\cdot(\mathbf{B}e_j),$$

where e_i are the standard basis vectors.

In particular, the coefficients of CC may be positively correlated even when those of $\mathbf{y}$ are uncorrelated, and their standard deviations are given by

$$\sigma(\mathrm{CC}(\mathbf{A}, p, T, \mathbf{y})_i) = \sigma||\mathbf{B}e_i||$$

Note that, by definition of $\mathbf{B}$, $\mathbf{B}e_i = \mathrm{CC}(\mathbf{A}, p, T, e_i)$, whose jth coefficients represents

the expected number of times node i is reached by the diffusion process, if seeded at node j.

By Proposition 3.5.1, we have the bound $\sigma(\text{CC}(\mathbf{A}, p, T, \mathbf{y})_i) = \sigma||\mathbf{B}e_i|| \leq \frac{\sigma}{1-p\lambda_1}$ if $p\lambda_1 < 1$.

3.5.3 Theoretical results of contextual centrality for Erdos-Renyi networks

In the case that $\mathbf{A}$ corresponds to an Erdos-Renyi graph $G(n, q)$, we have further theoretical results, in line with the results of Banerjee et al.[24] on diffusion centrality. As is standard for Erdos-Renyi graphs, we assume each edge has independent probability q of being present in the graph, where q is a function of n, the number of nodes. Assume that qn grows such that $log(n) \leq qn \leq \sqrt{n}$. We also assume that T and p are functions of n, and let $\mathbf{y}$ denote the vector (depending on n) consisting of $y_1, \ldots, y_n$ for some infinite sequence $\{y_i\}$. We suppress all dependency on n for ease of notation. We further assume that the mean $\overline{\mathbf{y}}$ has a limit $\overline{\mathbf{y}}$ as n approaches infinity, which is reasonable by the law of large numbers if the y_i are sampled from a random variable. With this background, we study the expected behavior of $E(\text{CC}(\mathbf{A}, p, T, \mathbf{y}))$.

Given two functions $f(n), g(n)$, we will say that f approaches g as n approaches infinity, if $lim_{n\to\infty}\frac{f(n)}{g(n)} = 1$. Then we have the following result.

Theorem 3.5.1. *Suppose $T = o(qn)$ and $\log(n) \leq qn \leq \sqrt{n}$. Then we can decompose $E(CC(\mathbf{A}, p, T, \mathbf{y})_i) = \overline{\mathbf{y}}E_1 + y_iE_2$, where E_1 and E_2 are functions of n, p, q, T but do not depend on $\mathbf{y}$ or i, such that*

a)E_1 approaches $\frac{1-(npq)^{T+1}}{1-(npq)}$.

b)$E_2 = o(E_1)$.

c)If $\overline{\mathbf{y}} \neq 0$, then $E(CC(\mathbf{A}, p, T, \mathbf{y}))$ approaches $\overline{\mathbf{y}}E(DC(\mathbf{A}, p, T, \mathbf{y}))$, where DC is diffusion centrality.

In other words, if $\overline{\mathbf{y}} \neq 0$, then the term E_1 dominates, so the expected CC is uniform all nodes (in the limit as n approaches infinity). Moreover, $\overline{\mathbf{y}}$ measures the magnitude of the diffusion as compared to DC, and the sign of $\overline{\mathbf{y}}$ determines the

expected sign of CC. In contrast, if $\overline{\mathbf{y}} = 0$, then CC equals E_2 so, on expectation, CC correlates perfectly with $\mathbf{y}$ itself. We note that in practice it is not likely for $\overline{\mathbf{y}}$ to equal 0. However, if $\overline{\mathbf{y}}$ is close to 0 and n is not too large, then the term E_2 could still be significant, indicating that the expected CC will be correlated with the nodal evaluation vector $\mathbf{y}$.

This result can also be related to the tradeoff in Eq. (3.6). As implied by the Theorem, as long as $\bar{y} \neq 0$, then expected CC approaches $\bar{Y}E_1$, which in turn approaches $\bar{y}E(\text{DC})$ as n approaches infinity. Thus the second term of the tradeoff in Eq. (3.6) dominates, on expectation.

We also note that careful analysis will show that $E_2 > 0$, but that is beyond the scope of the present project.

Theorem 3.5.2. *If $p\lambda_1 \geq (1+\epsilon)$ for some $\epsilon > 0$, then $T = \frac{log(n)}{log(npq)}$ is a threshold for viral spread if $\overline{\mathbf{y}} \neq 0$, in the sense that*

a)If $T \leq (1-\epsilon)\frac{log(n)}{log(npq)}$ for some $\epsilon > 0$, then $E(CC(\mathbf{A}, p, T, \mathbf{y})_i) = o(n)$ for all i.

b)If, on the other hand, $T \geq (1+\epsilon)(\frac{log(pnq)}{log(n)})$, then $E(CC(\mathbf{A}, p, T, \mathbf{y})_i) = \Omega(n)$ for all i.

Note that the threshold $T = \frac{log(n)}{log(npq)}$ given above is equal to $\frac{log(n)}{log(pE(\lambda_1))}$, since $E(\lambda_1) = nq$. We also note that the expected diameter of the Erdos-Reyni graph is $\frac{log(n)}{log(nq)}$, which is strictly smaller than the threshold given above.

To prove these theorems, we analyze $E(\mathbf{A}^t)$ for any t. Note that $E(\mathbf{A}^t)_{ij}$ is the weighted sum of all paths of length t from i to j, with each path π weighted by $q^{d(\pi)}$, where $d(\pi)$ is the number of distinct edges along the path π. Note that by symmetry, the off-diagonal entries of $E(\mathbf{A}^t)$ are all the same, as are its diagonal entries; however, the diagonal entries are not necessarily equal to the off-diagonal ones.

We first prove the following lemma to aid our analysis.

Let i, j, k be distinct numbers ranging from 1 to n. Let $Z_{ij,k}(t)$ be the subset of paths of length t from i to j which visit vertex k at some point. Let $z_{ij,k}(t)$ be its weighted sum $\sum_{\pi \in Z_{ij,k}(t)} q^{d(\pi)}$. Then $z_{ij,k} \leq \frac{t-1}{n-2} E(\mathbf{A}^t_{ij})$.

Proof. There are $(t-1)$ possible indices to place the vertex k. For each fixed index,

the weighted sum of all paths with vertex k at that index is $\leq \frac{1}{n-2}E(\mathbf{A}_{ij}^t)$, which follows by symmetry with respect to the $n-2$ possible choices of k. Combining these factors yields the desired bound. $\square$

We now move on to the estimates of $E(\mathbf{A}_{ij}^t)$.

For the purposes of this lemma assume that $\frac{t}{nq} \leq r < \frac{1}{4}$ for some r. Then we have

a) $(1-2r)\frac{(nq)^t}{n} \leq E(\mathbf{A}^t)_{ij} \leq (\frac{1}{1-4r})\frac{(nq)^t}{n}$, if $i \neq j$ or if $i = j$ and t is odd.

b) $(1-2r)\frac{(nq)^t}{n} \leq E(\mathbf{A}^t)_{ii} \leq (\frac{1}{1-4r})(\frac{(nq)^t}{n} + (2nq)^{t/2})$ if t is even.

Proof. Let us represent a path by the sequence of the vertices it visits. A path π of length t from i to j is represented as $iv_1v_2\cdots v_{t-1}j$, where i and j will also be labeled v_0 and v_t, respectively.

We begin by proving the lower bounds. We have $E(A_{ij}^t) \geq (n-2)^{t-1}q^t$. Indeed, there are more than $(n-2)^{t-1}$ legitimate paths in $X_{ij}(t)$ (under the constraint of no self-edges), and each one has at most t distinct edges. Now, $(n-2)^{t-1} \geq n^{t-1} - 2t(n)^{t-2} = n^{t-1}(1-\frac{2t}{n}) \geq (1-2r)n^{t-1}$ since $\frac{t}{n} \leq \frac{t}{qn} \leq r$.

Next, we calculate the upper bounds. Suppose that $t \geq 1$. Let $Y_{ij}(t) \subset X_{ij}(t)$ consist of those paths in which edges are never repeated immediately, that is, $v_l \neq v_{l+2}$ for any index l. Let $y_{ij}(t) = \sum_{\pi \in Y_{ij}(t)} q^{d(\pi)}$ be its weighted sum. We further partition $Y_{ij}(t)$ as follows. For each $k = 1, \ldots, (t-1)$, let $Y_{ij,k}(t) \subset Y_{ij}(t)$ be the subset of those paths for which k is the smallest index such that the edge $v_{k-1}v_k$ is not revisited later in the path, and $v_k \neq j$. Then $Y_{ij}(t) = \bigsqcup_{k=0}^{t-1} Y_{ij,k}(t)$. Let $y_{ij,k}(t)$ be the weighted sum of $Y_{ij,k}(t)$. Also, let $y_{\text{diff}}(t)$ and $y_{\text{same}}(t)$ denote the values of $y_{ij}(t)$ in the cases $i \neq j$ and $i = j$, respectively.

We will use the following properties for paths $\pi \in Y_{ij,k}(t)$. Given π, let $\pi' \in Y_{v_kj}(t-k)$ be the truncated path $v_k, \ldots, v_{t-1}, j$. We note that π has at least one edge that π' does not, namely $v_{k-1}v_k$, by definition of k. Thus $d(\pi) \geq d(\pi') + 1$. Furthermore, we note that every node $v_1, \ldots, v_{k-1}$ must be present in π'. Indeed, for each such vertex v, consider the greatest index l such that $v_l = v$. If $l < k$, then, by definition of k, that means either $v_l = j$, in which case it appears in π', or the edge $v_{l-1}v_l$ reappears later in the path. By assumption that $\pi \in Y_{ij}(t)$, this edge cannot

be repeated immediately; hence $v = v_l$ itself must reappear later, contradicting the description of the index l. So, $l \geq k$, that is, v indeed appears in π'.

These observations imply the following bound:

$$y_{ij,k}(t) \leq t^{k-1} nq y_{\text{diff}}(t-k) \tag{3.9}$$

Indeed, to specify a path in $Y_{ij,k}(t)$, we first choose v_k from among $\leq n$ possibilities. Then we choose the truncated path π' as described above from $Y_{v_k j}(t-k)$, whose weighted sum is $y_{v_k,j}(t-k)$. Then we choose the $k-1$ vertices $v_1, \ldots, v_{k-1}$. Each of them is repeated in π', hence may be chosen from among the $\leq t$ vertices of π'. Finally, since $d(\pi) \geq d(\pi') + 1$, we introduce the additional factor of q.

Now we focus on the case that $i \neq j$.

If $k \geq 1$, we can improve our bound further. Notice that, since $k > 1$, the starting vertex i must appear in the path π'. So, either $i = v_k$, or $i \neq v_k$. In the former case, we can eliminate a factor of n from (3.9), and in the latter case, we can introduce a factor of $\frac{t}{n}$ into (3.9), by Lemma 3.5.3 (Note the Lemma applies since i, j, and v_k are assumed distinct). We thus obtain the tighter bound

$$y_{ij,k}(t) \leq t^k q \cdot y_{\text{diff}}(t-k) \tag{3.10}$$

Now we can prove by induction that $y_{\text{diff}}(t) \leq (nq+2)^t$. Indeed, under this inductive hypothesis, the above bounds yield

$$y_{ij,1}(t) \leq nq(nq+2)^{t-1}$$

and

$$\sum_{k=2}^{t-1} y_{ij,k}(t) \leq \sum_{k=2}^{\infty} t^q k (nq+2)^{t-k} \leq t^2 q (nq+2)^{t-2} \frac{1}{1-r}$$
$$\leq 2(nq+2)^{t-1}$$

Where we used the fact that $(t^2) \leq (nq)^2 \leq n$ and $\frac{1}{1-r} \leq 2$. Combining these bounds

together we obtain, as desired, that

$$y_{ij}(t) = y_{ij,1}(t) + \sum_{k=2}^{t-1} y_{ij,2}(t) \leq (nq+2)(nq+2)^{t-1} = (nq+2)^t$$

Next, we plug in this bound for $y_{\text{diff}}(t)$ into (3.9), to obtain a bound for $y_{ij}(t)$ (even if $i = j$). We have

$$y_{ij} \leq \sum_{k=1}^{\infty} t^{k-1}(nq+2)^{t-k+1} \leq \frac{1}{1-r}(nq+2)^t$$

Now, it is convenient to further bound $(nq+2)^t \leq \frac{1}{1-2r}(nq)^t$. Indeed, $(nq+2)^t = \sum_{k=0}^{t} \binom{t}{k} 2^k n^{t-k} \leq \sum_{k=0}^{\infty} (2t)^k (nq)^{t-k}$, which is a geometric series with ratio $\frac{2t}{nq} \leq 2r$, bounded by $\frac{1}{1-2r}(nq)^t$.

Hence, we obtain the following bound on $y_{ij}(t)$:

$$y_{ij}(t) \leq \frac{1}{(1-r)(1-2r)}(nq)^t \leq \frac{1}{1-3r}(nq)^t \tag{3.11}$$

We emphasize that this inequality holds only if $t \geq 1$. Finally, we extend our analysis from the $Y_{ij}(t)$ to all paths. Any arbitrary path from i to j of length t may be obtained by starting with a path in $Y_{ij}(t-2m)$, for some $0 \leq m \leq t/2$ and performing a sequence of m insertions, replacing a vertex v with vwv instead, for some vertex w. We obtain the bound

$$E(\mathbf{A}_{ij}^t) \leq \sum_{m=0}^{\lfloor t/2 \rfloor} y_{ij}(t-2m) \cdot (2nq)^m \tag{3.12}$$

Indeed, for each insertion operation, there are two cases: either the inserted vertex w is already present in the path, so it can be chosen from among $\leq t$ vertices; or it is not already present, in which case it can be chosen from among $\leq n$ vertices and introduces a new edge, for an additional factor of q. Combining the two possibilities, each insertion operation introduces a factor of $(t + nq) \leq 2nq$.

To evaluate this sum, we need to consider the two cases outlined in the statement

of this lemma.

a) Suppose that either $i \neq j$, or $i = j$ and t is odd. In this case, note that the bound (3.11) can be applied to each $y_{ij}(t-2m)$, since if t is odd, then $t - 2m \geq 1$; and if $i \neq j$, we have $y_{ij}(0) = 0$ regardless. Combining these bounds with (3.12), we obtain

$$E(\mathbf{A}_{ij}^t) \leq \frac{1}{(1-3r)} \sum_{m=0}^{\infty} \frac{1}{n} 2^m (nq)^{t-m} \leq \frac{1}{1-4r} \frac{1}{n} (nq)^t$$

by a geometric series with ratio $\frac{2}{nq} < r$. This completes the proof of part a) of the lemma.

b) Now suppose that $i = j$ and t is even. The sum in (3.12) can be analyzed in the same way as in a), but with an extra term of $(2qn)^{t/2}$ corresponding to the case $m = \frac{t}{2}$. □

We are now ready to prove Theorem 3.5.1.

Proof. By definition, $E(\mathrm{CC}(\mathbf{A}, p, T, \mathbf{y})) = E(\sum_{t=0}^{T} p^t \mathbf{A}^t \mathbf{y})$. By linearity of expectation, this equals $\sum_{t=0}^{T} p^t E(\mathbf{A}^t)\mathbf{y}$. Now, for each t and each i, we have $(E(\mathbf{A}^t)\mathbf{y})_i = \sum_{j=0}^{n} y_j p^t E(\mathbf{A}_{ij}^t)$. By separating the terms with $i = j$ from the terms with $i \neq j$, this equals $n\overline{\mathbf{y}} p^t E(\mathbf{A}_{\mathrm{diff}}^t) + y_i p^t \cdot (E(\mathbf{A}_{\mathrm{same}}^t) - E(\mathbf{A}_{\mathrm{diff}}^t))$, so we can write

$$E(\mathrm{CC}(\mathbf{A}, p, T, \mathbf{y})_i) = \overline{\mathbf{y}} E_1 + y_i E_2$$

where

$$E_1 = n \sum_{t=0}^{T} p^t E(\mathbf{A}_{\mathrm{diff}}^t)$$

and

$$E_2 = \sum_{t=0}^{T} p^t \cdot (E(\mathbf{A}_{\mathrm{same}}^t) - E(\mathbf{A}_{\mathrm{diff}}^t))$$

a) By Lemma 3.5.3, we know that E_1 can be bounded

$$(1-2r) \sum_{t=0}^{T} (npq)^t \leq E_1 \leq \frac{1}{1-4r} \sum_{t=0}^{T} (npq)^t$$

where $r = \frac{T}{nq}$. Since we assume this ratio approaches 0, these bounds imply that

indeed E_1 approaches $\sum_{t=0}^{T}(npq)^t = \frac{(npq)^{T+1}}{1-npq}$ as n tends to infinity.

b) Next, we show that $E_2 = o(E_1)$. Indeed, we again use Lemma 3.5.3. We have

$$|E_2| \leq \sum_{t=0}^{T} p^t \cdot (E(\mathbf{A}^t_{\text{same}}) + E(\mathbf{A}^t_{\text{diff}}))$$

$$\leq \frac{1}{1-4r}(\sum_{t=0}^{T}(p^t(2nq)^{t/2}) + \frac{(npq)^t}{n})$$

so the result follows since both terms $p^t(2nq)^{t/2}$ and $\frac{(npq)^t}{n}$ are lower-order than $(npq)^t$.

c) Diffusion centrality is a special case of contextual centrality in which $\mathbf{y} = \mathbf{1}$, which has mean 1. The result follows by part a), together with the fact that E_1 dominates over E_2 whenever $\overline{\mathbf{y}} \neq 0$ by part b). □

Next, we prove Theorem 3.5.2.

Proof. Suppose $\overline{\mathbf{y}} \neq 0$ and $pE(\lambda_1) \geq (1+\epsilon)$ and that. For Erdos-Renyi graphs, $E(\lambda_1) = nq$, so $pnq \geq (1+\epsilon)$. In this case, it follows from Theorem 3.5.1 that $E(\text{CC})_i$ approaches $\overline{\mathbf{y}}(pnq)^T$. If $T \leq (1-\epsilon)\frac{log(n)}{log(npq)}$ for some $\epsilon > 0$, then $log(|E(\text{CC})_i|) \leq C + log(|(pnq)^T|)$ for some constant C, which equals $C + Tlog(pnq) \leq C + (1-\epsilon)log(n)$ so $T = O(n^{1-\epsilon}) = o(n)$. The other direction follows similarly. □

3.5.4 The relationship between contextual centrality and other centrality measures

Degree, eigenvector, Katz, diffusion, and contextual centrality can all be expressed as specific cases of a simple recurrence relation with an intuitive explanation. Roughly speaking, a node's importance in a network can be broken down into two parts: its influence on other nodes in the network through its neighbors, and its individual contribution to the cascade payoff.

Let $\mathbf{c}_t$ be the importance (i.e., centrality) of all nodes in the network at time step t. One way to capture the notion of each node's influence on other nodes in the network is through $\mathbf{A}\mathbf{c}_{t-1}$, where $\mathbf{A}$ is the adjacency matrix of the network. This term

effectively sums up the importance of the neighbors of each node. With this in mind, we can express $\mathbf{c}_t$ as,

$$\mathbf{c}_t = \alpha \mathbf{A}\mathbf{c}_{t-1} + \boldsymbol{\beta}, \tag{3.13}$$

where α is a constant and $\boldsymbol{\beta}$ is the individual contribution of each node in the network. It is, of course, possible to parameterize α, $\boldsymbol{\beta}$, or $\mathbf{A}$ by t as well, but for simplicity let us assume they remain constant. Expanding this recurrence, we get

$$\mathbf{c}_t = (\alpha \mathbf{A})^t \mathbf{c}_0 + \sum_{i=0}^{t-1} (\alpha \mathbf{A})^i \boldsymbol{\beta}. \tag{3.14}$$

Now if we substitute $\alpha = p$, $\boldsymbol{\beta} = \mathbf{y}$, and $\mathbf{c}_0 = \mathbf{y}$, then $\mathbf{c}_T$ is exactly equal to CC. Substitutions can be done for all the centrality measures discussed above and are summarized in Table 3.1.

Table 3.1: **Centrality measures defined by $\mathbf{c}_t = \alpha \mathbf{A}\mathbf{c}_{t-1} + \boldsymbol{\beta}$.**

Centrality	α	$\boldsymbol{\beta}$	$\mathbf{c}_0$	t
Degree	1	$\mathbf{0}$	$\mathbf{1}$	1
Eigenvector	1	$\mathbf{0}$	$\mathbf{1}$	∞
Katz	$< \frac{1}{\lambda_1}$	$\mathbf{1}$	$\mathbf{1}$	∞
Diffusion	p	$\mathbf{1}$	$\mathbf{1}$	T
Contextual	p	$\mathbf{y}$	$\mathbf{y}$	T

Contextual centrality is developed upon and generalizes diffusion centrality, but there are two important differences. First, all nodes passed through by the random walk contribute positively and homogeneously in diffusion centrality, while the main advantage of contextual centrality is allowing for the heterogeneous contributions. Second, the random walk of contextual centrality starts from the chosen seed, while that of diffusion centrality starts from the neighbors of the chosen seed. Under the condition that $\overline{\mathbf{y}}$ is positive and constant for all entries, contextual centrality inherits the nice nesting properties of diffusion centrality, which encompasses and spans the gap between degree centrality, eigenvector centrality, and Katz centrality. In particular, CC is proportional to degree centrality when $T = 1$, proportional to eigenvector

centrality as $T \to \infty$ when $p \geq \lambda_1^{-1}$, and proportional to Katz centrality when $T = \infty$ and $p < \lambda_1^{-1}$. Proof can be found in Banerjee et al. [27].

Contextual centrality is also similar to Katz centrality, but we highlight two crucial differences. First, contextual centrality is more general in that p can be larger than λ_1^{-1} and provides essential insights into this region. Second, we allow T to vary according to the specific setting, while in Katz centrality, the diffusion period T is infinite. T carries important implications. For the product that is effective in a short period, such as a coupon that will expire within a day, T is relatively small compared with the diffusion of a new phone, which will be on the market for much longer.

3.5.5 Relationship between approximated cascade payoff and cascade payoff

Contextual centrality aims to maximize objective (3.4), which provides an approximation to cascade payoff, as in objective (3.3), by an independent cascade model. In Fig. 3-8, we analyze the Spearman and Pearson correlation between the two concerning different spreadability. Both correlation measures decrease as spreadability increases from 0 to 1 and increase afterward. In the bulk part, Spearman's correlation between the two is higher than the Pearson correlation and is around 0.9 or higher. Note that $p\lambda_1 = 1$ is the phase transition in network contagion with the Susceptible-Infected (SI) model and is known as the epidemics threshold[60]. This may explain why we see a different behavior close to $p\lambda_1 = 1$.

3.5.6 Game-theoretic interpretation of contextual centrality with local interactions

Ballester et al. are the first to provide a behavioral foundation for centrality, in particular, Katz-Bonacich centrality, using a complementary linear-quadratic-form network game [20]. They found that one's network position can fully explain the Nash equilibrium in such network games. Similarly, we show that when the spreadability is smaller than one, and agents can interact for an infinite time period, their activity

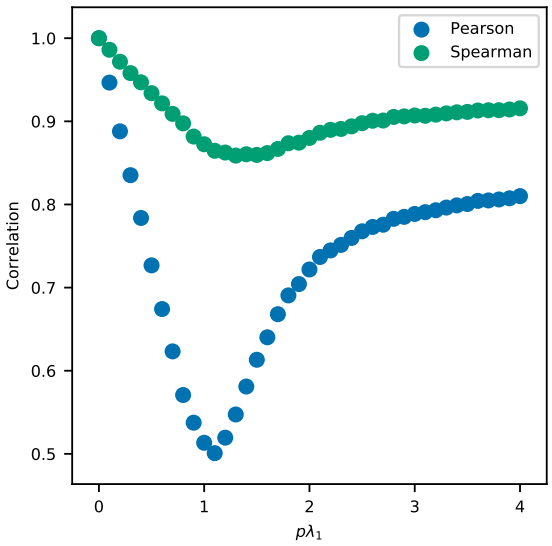


Figure 3-8: **Relationship between approximated cascade payoff and cascade payoff.** The y-axis and x-axis display the correlation and the spreadability ($p\lambda_1$) respectively. Pearson and Spearman's correlation are shown in blue and orange color respectively.

levels can be explained by both their structural positions, as well as their marginal benefits of actions (which corresponds to nodes' contributions).

In the setup of Ballester et al., agents choose actions optimally in response to their neighbors [20]. The quadratic functional form implies that the utility of individual i, (u_i), is quadratic in i's action level (a_i), are dependent on i's neighbors' effort, and has a homogeneous marginal benefit α across the population, $u_i = \alpha a_i - \frac{1}{2}a_i^2 + \beta \sum_{j=1}^{n} A_{ij}a_i a_j$, where α is a scalar and $\alpha > 0$. Taking the first-order condition, it is easy to prove that the strategy in Nash equilibrium is $\mathbf{a} = \left(\mathbf{I} - \beta\mathbf{A}\right)^{-1}\alpha \propto c_{\text{Katz}}$, which is proportional to the Katz centrality.

In the previous setup, Ballester et al. assume that the marginal benefit is homogeneous and positive. We relax this constraint, allowing it to vary across individuals (y_i) with and can take on negative values. With this, suppose agent i chooses an action (a_i) according to the following utility function,

$$u_i = a_i y_i - \frac{1}{2}a_i^2 + \beta \sum_{j=1}^{n} A_{ij}a_i a_j. \tag{3.15}$$

With this variant, the equilibrium strategy becomes,

$$\mathbf{a} = \left(\mathbf{I} - \beta\mathbf{A}\right)^{-1}\mathbf{y}. \tag{3.16}$$

Eq. (3.16) has the exact same form as CC when $T \to \infty$, $\beta\lambda_1 < 1$ and $\beta = p$. Hence, we see that contextual centrality approximates agents' equilibrium actions with heterogeneous marginal utilities in this condition.

3.5.7 Differences between contextual centrality and centrality measures developed on weighted networks

There have been some studies that generalize centrality measures to weighted or signed networks. They focus on settings where edge weights represent the strength or the trustiness (a friend or a foe) of the social relationships. The network information captured by these centrality measures can be regarded as a special case of a weighted version of contextual centrality, where $\mathbf{A}$ is a weighted matrix, $p = 1$ and $\overline{\mathbf{y}} = \mathbf{1}$. Weights on network links emphasize social relationships but do not capture the heterogeneous contributions of the nodes - exogenous to the network structure - directly to the cascade payoff. Weighted links and weighted nodes characterize different network dynamics and diffusion objectives. Let us provide a simple illustrative example to explain the differences better. Imagine a network with two disconnected communities, where one component consists of positive links, and the other consists of negative links. Centrality developed on the weighted or signed network will rank the most-connected node as the top in the community with positive edges (i.e., individuals who all trust one another). However, for a particular marketing campaign, if all individuals in the positive community do not like the product, seeding any individuals in the positive community will hurt the campaign.

For readers' reference, we provide an overview of centrality measures on weighted and signed networks. There are two main strands of work in this literature. First, some studies define new notions of the shortest path that take the weights of the links into account. There are multiple types of modifications: (1) take the inverse of the

tie strengths as the shortest path lengths [155, 51], (2) using a tuning parameter to trade-off tie strengths and the number of ties [157], (3) adding a temporal aspect to links to minimize the temporal latency [199]. With these new notions, researchers extend existing path-based centrality measures. [155] and [51] extended closeness centrality and betweenness centrality to define the shortest path algorithm to be the least costly path with cost depending solely on tie weights. Opsahl proposes a centrality measure with a generalized degree and shorted paths computation by adding a tuning parameter on tie strengths. Another strand of studies focused on the flow and diffusion processes [157]. Kunegis et al. develop a signed centrality measure using the left eigenvector of the signed network as a generalization of the eigenvector centrality with weighted edges[126]. Other studies develop algorithm-based ranking methods, extending PageRank or HITS. Shahriari and Jalili compute the difference between the scores using PageRank or HITS algorithms for networks consisting of positive and negative links, respectively, as the new measure [171].

3.6 Additional results for empirical analysis

3.6.1 Predictive power of contextual centrality in eventual adoptions

Here we include the additional results to examine the robustness of the predictive power of contextual centrality in the eventual adoption outcomes similar to Fig. 3-1. We extend the linear regression models to (1) without controlling for village size (as shown in Fig. 3-9), and (2) with additional controls (as shown in Fig. 3-10).

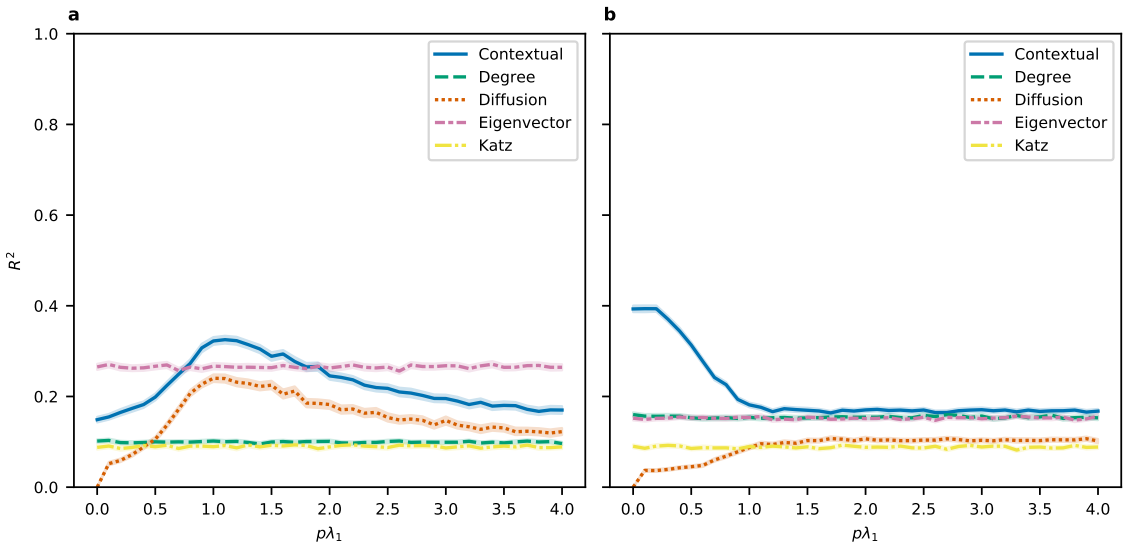


Figure 3-9: **Predictive power of contextual centrality without any controls for (a) microfinance and (b) weather insurance.** The y-axis shows the 95% confidence interval of R^2 computed from 1000 bootstrap samples from ordinary least squares regressions controlling for village size. The x-axis shows varying values for $p\lambda_1$, which influences only diffusion centrality and contextual centrality.

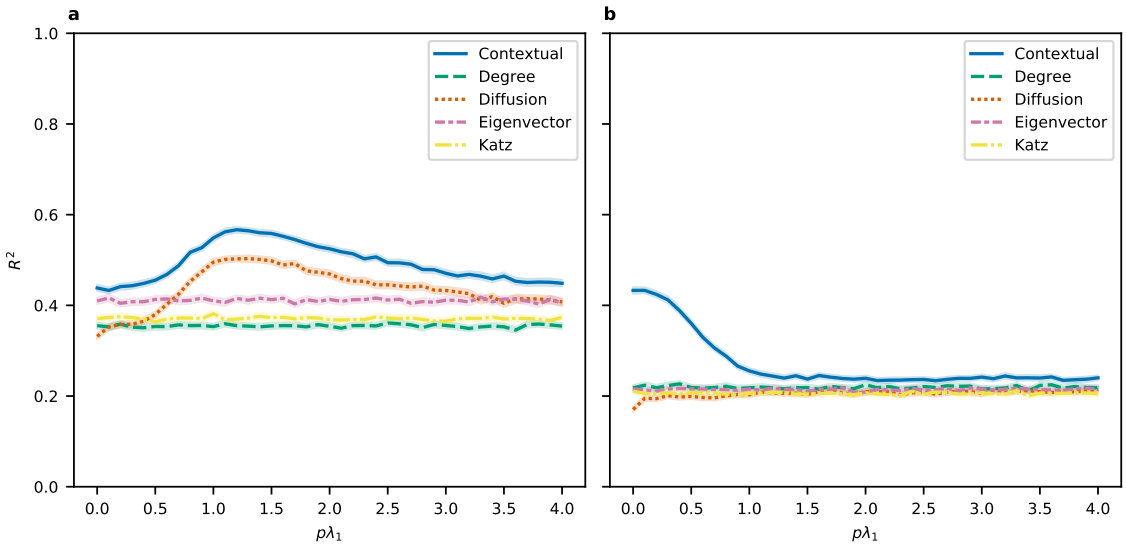


Figure 3-10: **Predictive power of contextual centrality with additional controls for (a) microfinance and (b) weather insurance.** For (a), we use village size, savings, self-help group participation, fraction of general caste members, and the fraction of village that is first-informed as done in [23]. For (b), we use village size, number of first-informed households, and fraction of village that is first-informed. The y-axis shows the 95% confidence interval of R^2 computed from 1000 bootstrap samples from ordinary least squares regressions controlling for village size. The x-axis shows varying values for $p\lambda_1$, which influences only diffusion centrality and contextual centrality.

3.6.2 Performance relative to other centrality measures on random networks

Here we show additional results corresponding to Fig. 3-3 and Fig. 3-4. From Fig. 3-11 to Fig. 3-19, we vary the standardized average contribution from -4 to 4. Note in all cases CC has an advantage over other seeding methods when the $p\lambda_1$ is small and loses some of this advantage as $p\lambda_1$ increases. The rate at which CC loses its advantage increases as the magnitude of the standardized average contribution increases. When $p\lambda_1$ is large CC performs comparably to other centrality measures, but in some cases still maintains an advantage.

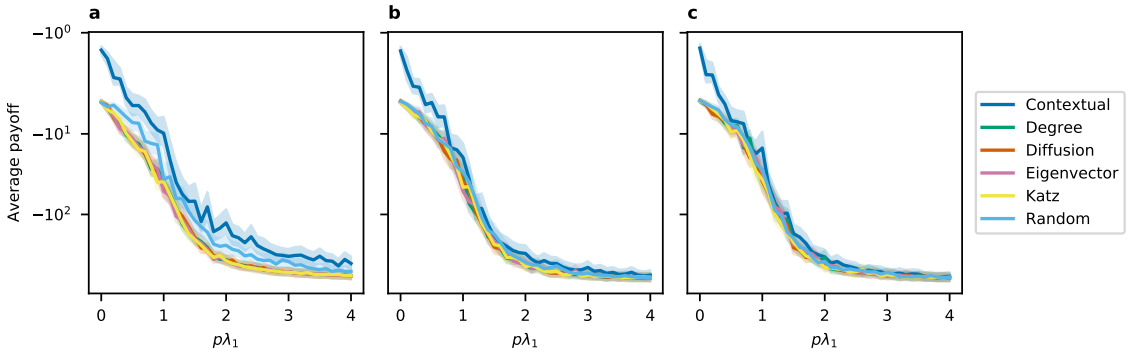


Figure 3-11: **Average payoffs with 95% confidence interval when standardized average contribution is -4** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

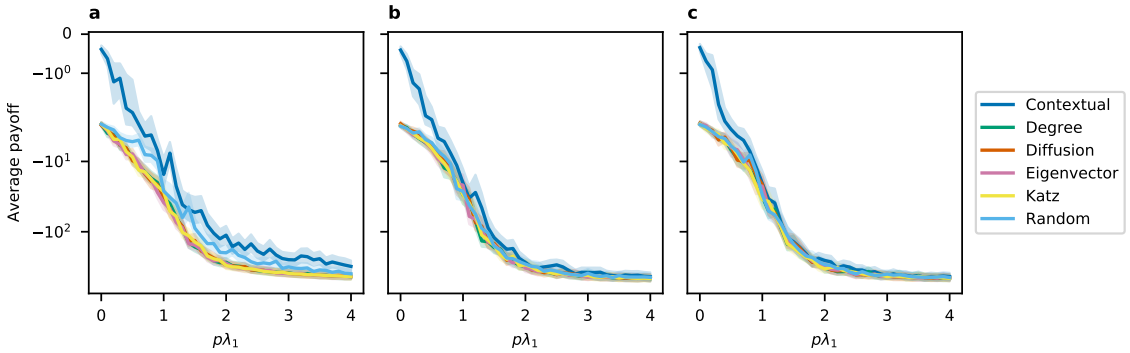


Figure 3-12: **Average payoffs with 95% confidence interval when standardized average contribution is -3** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

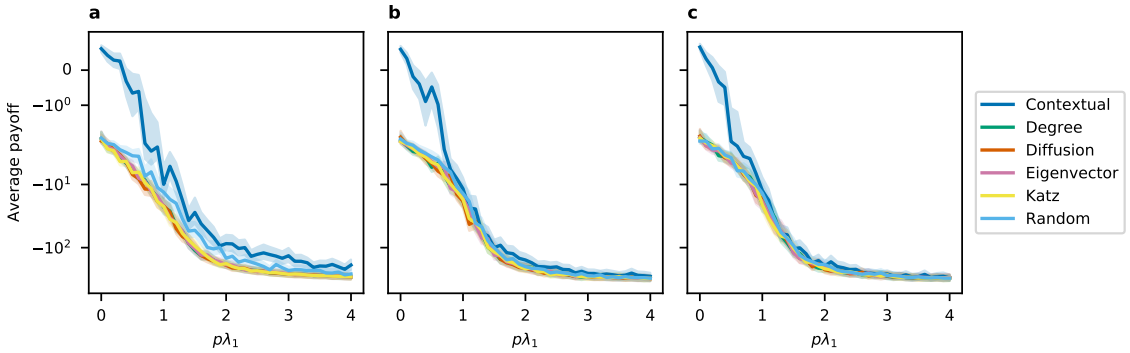


Figure 3-13: **Average payoffs with 95% confidence interval when standardized average contribution is -2** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

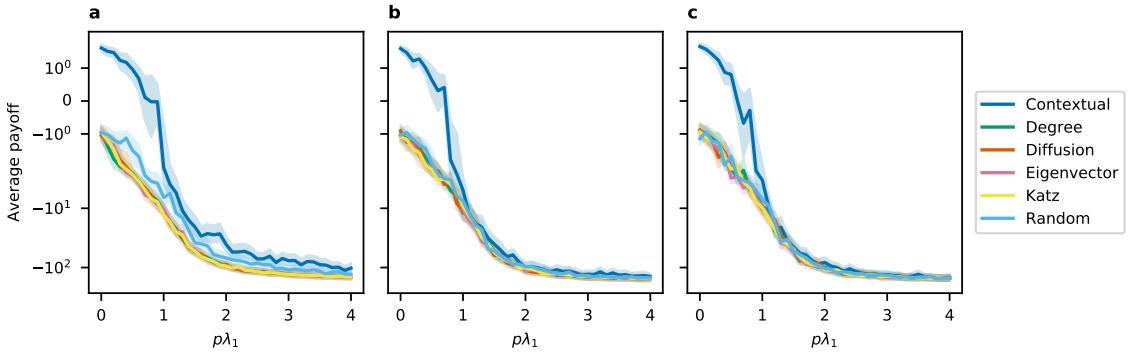


Figure 3-14: **Average payoffs with 95% confidence interval when standardized average contribution is -1** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

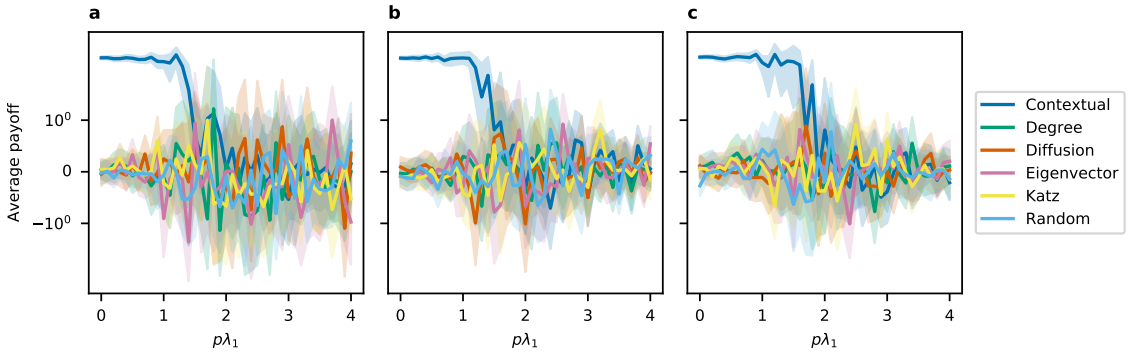


Figure 3-15: **Average payoffs with 95% confidence interval when standardized average contribution is 0** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

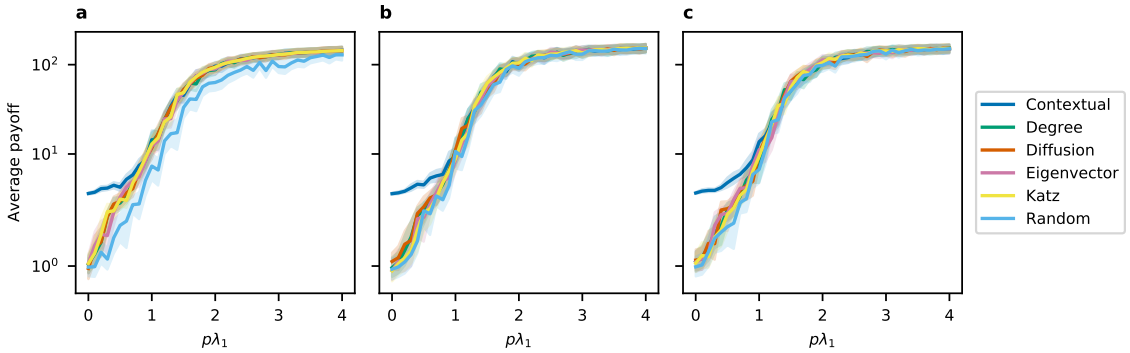


Figure 3-16: **Average payoffs with 95% confidence interval when standardized average contribution is 1** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

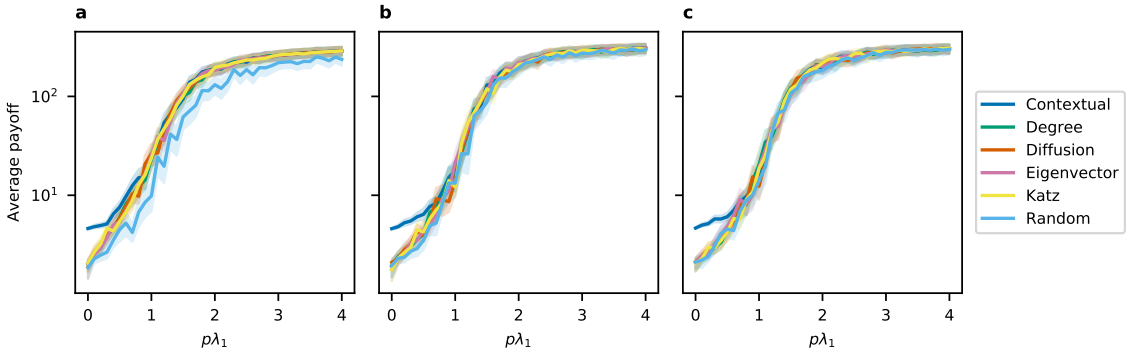


Figure 3-17: **Average payoffs with 95% confidence interval when standardized average contribution is 2** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

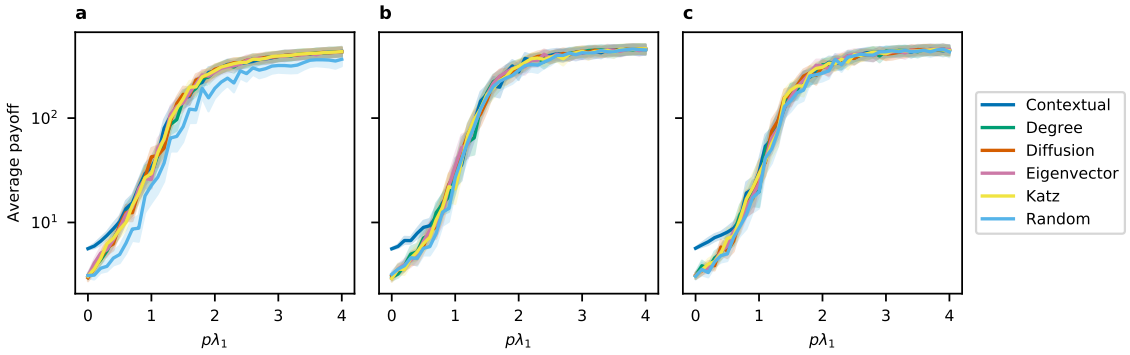


Figure 3-18: **Average payoffs with 95% confidence interval when standardized average contribution is 3** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

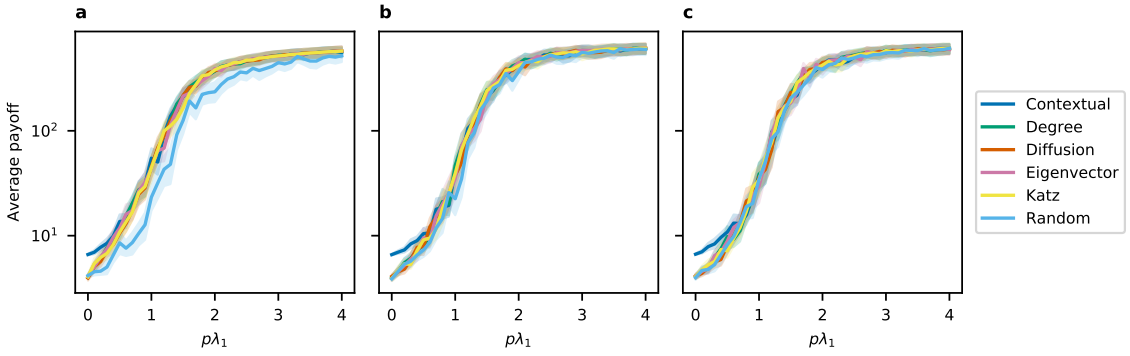


Figure 3-19: **Average payoffs with 95% confidence interval when standardized average contribution is 4** for (a) Barabasi-Albert, (b) Erdos-Renyi, and (c) Watts-Strogatz models.

3.6.3 Average approximated cascade payoff for contextual centrality and the variations of other centrality measures

Here we present the average approximated cascade payoff for contextual centrality and the variations of other centrality measures, including degree centrality (as shown in Fig. 3-20), diffusion centrality (as shown in Fig. 3-21), and Katz centrality (as shown in Fig. 3-22). Note that the approximation does not hold for degree centrality when $p\lambda_1 > 1$ and T is large. However, scaling degree centrality with primary contribution still improves the performance, so we present it here.

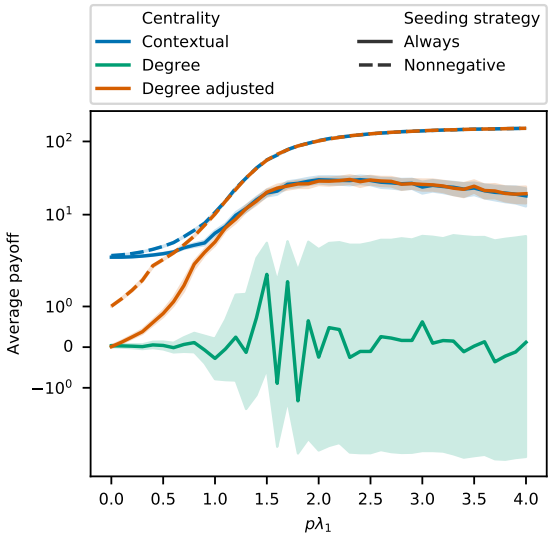


Figure 3-20: **Average cascade payoff for variations of contextual centrality and degree centrality.**

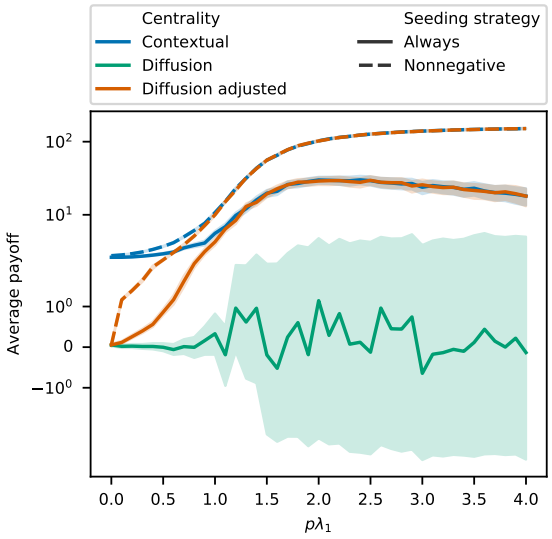


Figure 3-21: **Average cascade payoff for variations of contextual centrality and diffusion centrality.**

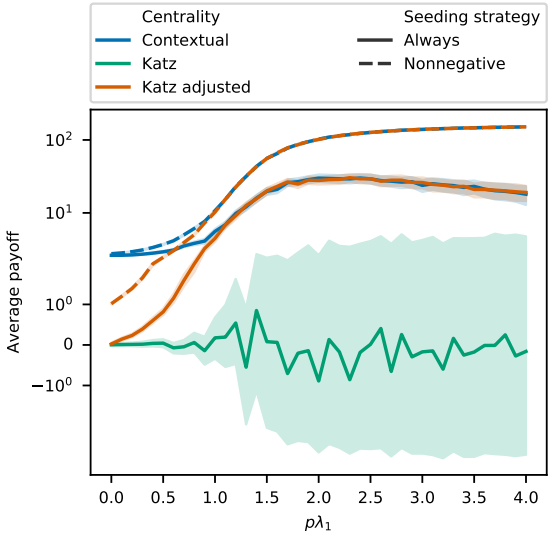


Figure 3-22: **Average cascade payoff for variations of contextual centrality and katz centrality.**

3.6.4 Comparison of seeding strategies when $\overline{\mathbf{y}}(\mathbf{U}_1^T\mathbf{y}) < 0$

Here we show the effect of using different seeding strategies on the average cascade payoff. For this plot, we generated 1000 random networks for each graph type (Barabasi-Albert, Erdos-Renyi, and Watts-Strogatz) as before with contributions sampled from a standard normal distribution for (a) continuous and sampled from $\{-1, 1\}$ with equal probability for (b) discrete. We redistributed the contributions to make the signs of $\overline{\mathbf{y}}$ and $\mathbf{U}_1^T\mathbf{y}$ differ if possible, and then filtered out results for which the signs did not differ. More specifically, if the average contribution was negative, the individual with the largest eigenvector centrality score was given the most positive contribution, the individual with the second-largest eigenvector centrality score was given the second most positive contribution, and so on. We used an analogous procedure if the average contribution was positive. Fig. 3-23 shows that seeding according to the contextual centrality score tends to perform the best as long a $p\lambda_1$ is not too large, after which seeding according to the average contribution performs the best. For small values $p\lambda_1$, seeding always performs as well as, if not better than, seeding according to contextual centrality. As suggested by Eq. (3.7), seeding according to the primary contribution yields similar results as seeding according to contextual centrality score as $p\lambda_1$ grows large.

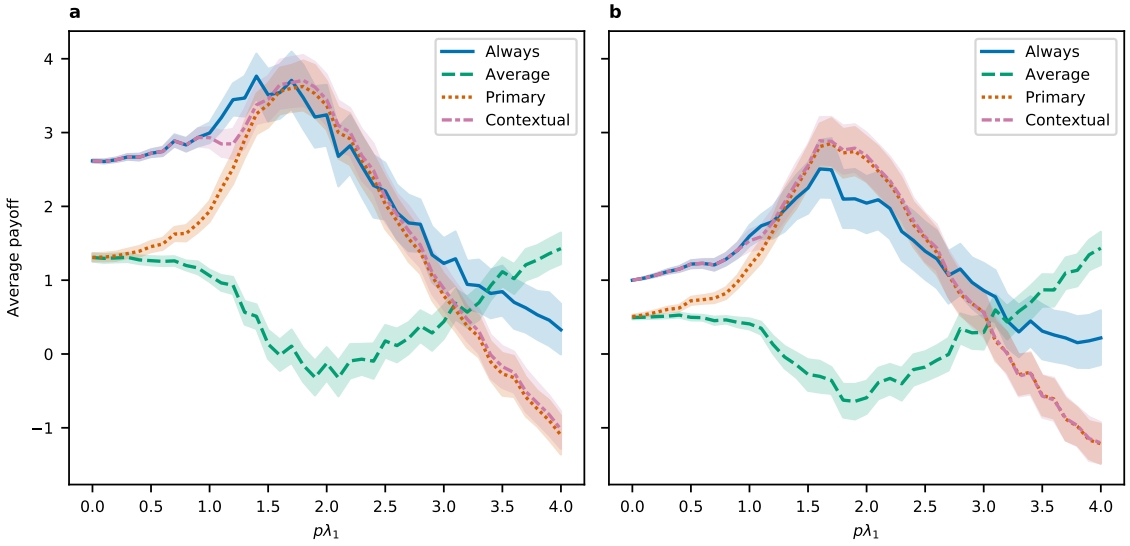


Figure 3-23: **Comparison of seeding strategies when $\overline{\mathbf{y}}(\mathbf{U}_1^T\mathbf{y}) < 0$ for (a) continuous and (b) discrete.**

Chapter 4

Recommender systems with heterogeneous information: A geometric deep learning approach

Many real-world data cover multiple aspects of consumer behaviors and business characteristics, creating opportunities for marketing companies to better understand the demand and preferences of customers with complementary information. However, to effectively combine data with multi-modal nature and complex structure is challenging. In this study, we propose a novel geometric deep learning framework for building effective recommender systems by predicting customers' preferences on businesses they have not yet rated. The proposed framework is capable of handling heterogeneous and auxiliary information on businesses and customers, and at the same time enforcing that only information relevant to the prediction task will be utilized. We compare the proposed framework with several baseline models in a prediction task using the Yelp open data set, where the improved performance of our method highlights the advantage of incorporating spatial, temporal, network, and other types of data in a principled manner. The proposed framework can be further applied to help make more informed marketing and managerial decisions in a variety of domains where the fusion of heterogeneous and structured information could be beneficial.[1]

[1]This work is joint with Rodrigo Ruiz, Xiaowen Dong, and Alex Pentland.

4.1 Introduction

Recommender systems are everywhere nowadays in the digital space. They significantly influence the information consumers receive and invisibly guide their behaviors [179, 133]. With overwhelming information flooding the Internet, recommender systems help reduce search cost [11] and uncertainties [34] of customers, as well as widen their interests and create commonalities among consumers [105]. Recommender systems also impact businesses in many ways. They affect sales [131] and help increase revenues and profits drastically, e.g., 60% of Netflix rentals and 35% of Amazon sales originate from recommender systems [105]. Moreover, for manufacturers, retailer's deployment of recommender systems influences their pricing strategy, i.e., the pricing competition may be either intensified or softened depending on the market characteristics [138]. [77] show that 94% of the e-commerce sites surveyed are now considering recommender systems as critical to current and future success, while 72% of companies attribute their business failure to the lack of knowledge on recommender systems. However, only 4% of organizations describe their website experience as "very" personalized. This highlights the importance of and opportunities in developing effective recommender systems in e-commerce and beyond.

The Big Data deluge, together with the rapid development of information technology, e.g., communication, mobile, and networking technologies [168], has proliferated business opportunities for companies and marketing departments. Customers' digital profiles have become unprecedentedly richer. For example, a digital platform may have the customers' location information, social connections, preferences and interests, and other demographic information. How to combine such heterogeneous information is, therefore, key to effective recommender systems. The core value of recommendation also increasingly lies in personalization [77, 179], which has become ubiquitous with a large amount of user information on the Internet. Both present significant challenges in developing sophisticated methods to combine various sources of information for a more personalized recommendation.

Existing recommender systems use statistical techniques to infer customers' prefer-

ences and products' characteristics to make recommendations that best match the two. There are three main approaches to inferring customers' preferences: (1) collaborative filtering [178], i.e., people who showed similar preferences in the past are likely to prefer the same product in the future; (2) matrix factorization [153], i.e., the rating matrix is factorized into a latent user matrix and a latent product matrix separately characterizing the hidden characteristics of the users and products, and recommendations are made based on the learned latent representations; (3) exploitation of social relationships [144, 180], i.e., people who are friends tend to like similar products.

These traditional approaches are, however, insufficient in addressing the aforementioned challenges, mainly in the following two aspects. First, they are mostly designed for handling single-source information rather than heterogeneous information from different sources, such as metadata attached to businesses or location information of users, which are however very helpful in developing effective recommender systems [206]. To incorporate heterogeneous information, recent approaches such as heterogeneous information network proposes to build meta-paths from users to products and, together with a corresponding similarity measurement along each path, make predictions on user preference [206]. However, since the heterogeneous information is processed independently, there may be information redundancy or loss when the paths are combined. Moreover, this method does not work well with noisy and sparse information, since it relies on explicit path reachability between the user and product [174]. Another general class of methods is multi-view learning, where heterogeneous information is usually stacked together, and each source of information contributes equally, or in a global manner, to the learning task. This is, however, an unrealistic assumption in many real-world scenarios, where different information sources may weigh differently [98].

Second, information about the social relationship between users, or similarity between products, may complement the collaborative filtering and matrix factorization approaches for more accurate recommendations. Taking social relationship as an example, however, due to the lack of information about the strength of user friendship, the friends of one customer are usually treated equally, i.e., they contribute equally

to the prediction of preference of that customer [144, 180]. This is not ideal in many scenarios, e.g., closer friends may have more similar tastes. shall we delete the noise part? I think it dilutes our main idea, and does not go very well with the motivation below. Moreover, there may exist noisy information where "friends" reported on online platforms do not correspond to friendship in the offline sense.

blue Let us use a simple example to illustrate the motivation of our method, as shown in Figure 4-1. Consider a set of individuals who are connected by two kinds of networks: professional and social. A platform wants to infer consumers' preferences on restaurants to make personalized recommendations. Let us assume people form a professional relationship on educational background, and social relationship based on their hobbies and how people spend their spare time. So only social network provides information about user preferences on restaurants. Suppose we only observe connections, but we cannot distinguish these two types of relationships. Now, if we also observe hobbies and how people spend their sparse time, that information might be useful in disentangling social relationships from professional relationships, which in term will have predictive power over restaurant preferences. This means that even when we cannot observe social and professional network separately, if we have some auxiliary information, we can utilize their predictive power over both networks and Yelp reviews to pick out the useful component of network, in this case, their latent social network, and purge the noise that are not predictive over review, in this case, the professional network.

In this section of my thesis, we adopt techniques developed in the emerging field of geometric deep learning [52], particularly the graph attention networks [192], and propose a novel framework for recommending businesses to users given heterogeneous information. Specifically, we consider a matrix completion problem, i.e., predict the missing entries in a partially observed user-business rating matrix and make recommendations to users on the businesses with high predicted ratings, as illustrated in Figure 4-2. To this end, we develop a geometric deep learning architecture to learn low-dimensional latent representations for both users and businesses for the recommendation. The core idea of our approaches is that these low-dimensional

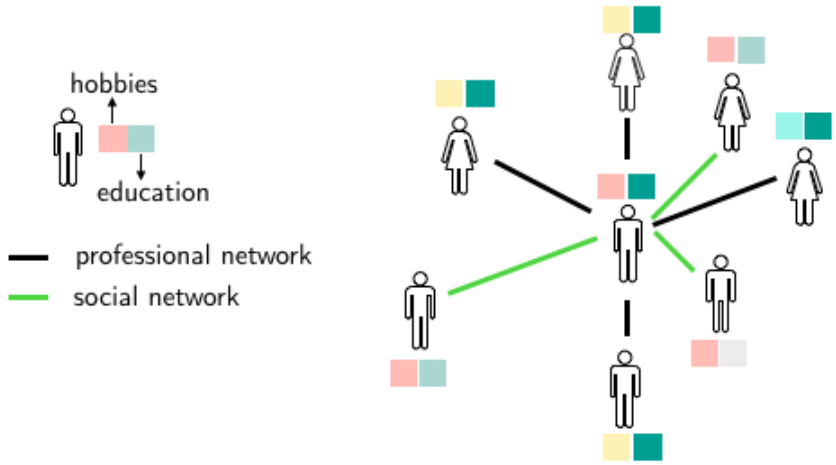


Figure 4-1: **Motivating example.** Each individual is characterized by hobbies and educational background. The black and green link correspond to professional and social networks, respectively.

representations are obtained by aggregating a diverse set of information from the neighbors of each user or business in a network, where the influences of the neighbors are weighted according to their relevance to the target user or business. In other words, a relevance score is assigned to one's neighbors which helps filter out noisy information and makes the resulting model both more effective and more interpretable. We demonstrate the meaningfulness of the proposed method and its superiority over some baseline approaches in a prediction task using an open data set from Yelp, an online business review platform, where we make use of the rich heterogeneous information about both users and businesses on the platform.

We summarize the main contributions of this project as follows:

1. We propose a framework to integrate heterogeneous information of different nature in a principled way. We merge spatial, temporal, network, and other types of data and selectively utilize the information that is predictive to the recommendation task. This is achieved by a novel geometric deep learning architecture, which is capable of handling both unstructured and structural information.

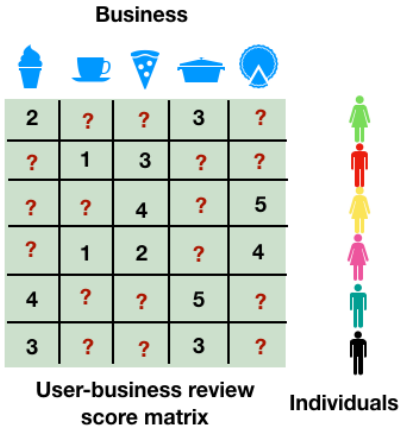


Figure 4-2: **Predicting customer preferences on businesses as a matrix completion task.**

2. Rather than assuming all user-user (or business-business) relationships to be the same, we weigh the neighbors of each user (or business) according to how relevant they are to the target one. This is achieved by the graph attention networks, which makes the model both more effective and more interpretable. Our method is especially helpful when the network connections are noisy or have missing information.

3. We perform several analyses to show that auxiliary information provides useful information in the task of predicting business ratings. We demonstrate the effectiveness of our method on the Yelp data set with rich information about businesses and users. We further show that the proposed method outperforms several benchmarks, which highlights the advantage of incorporating heterogeneous sources of information with respective importance. Furthermore, analysis of the latent business representations demonstrates that our method can effectively extract interpretable patterns about the characteristics of businesses.

4. The proposed framework has far-reaching marketing and managerial implications, especially for online and offline platforms that contain rich, heterogeneous, and in particular relational (network) information about customers or businesses. Learning preferences of customers or characteristics of businesses have a wide

range of applications in real-world business problems, including not only recommendation but also customer stratification, business grouping, and prediction and inference of user or business relationship.

The rest of this section of my thesis is organized as follows. We review relevant literature in Section 4.2. We describe the data in Section 4.3 and present some basic analyses. We then formulate the problem in Section A.3 and describe the proposed framework in detail. We present experimental results and analyze the patterns of the learned latent representations in Section 4.5, and conclude this chapter in Section A.5.

4.2 Literature review

4.2.1 Recommender systems

Personalized recommendations have become commonplace due to the widespread adoption of recommender systems. Major companies including Amazon, Facebook, Netflix, Pandora, and Yelp provide users with recommendations on a variety of products including friends, movies, songs, and restaurants [111]. There are two main approaches to recommender systems: collaborative filtering and content-based filtering [42]. Collaborative filtering recommends items based on the interests of users with similar ratings, while Content-based filtering recommends items that are similar to other items in which the user has expressed interest. Each of these approaches has its advantages and disadvantages. On one hand, collaborative filtering systems can glean information from multiple users/items without explicit knowledge of them. However, these systems have difficulty providing accurate recommendations when users/items have few ratings. Moreover, even the most active users have rated a tiny fraction of the items. These issues are known as the cold start and sparsity problems [42], respectively. On the other hand, content-based filtering systems do not face these problems, but they require explicit knowledge of the users/items that is often difficult to extract reliably, and this can adversely affect recommendations. Moreover, these systems suffer from a lack of diversity in their recommendations. These issues are known as

the limited content analysis and overspecialization problems [42], respectively. Hybrid approaches incorporate ideas from both collaborative and content-based filtering in an attempt to solve the problems of both approaches [54].

Since recently, recommender systems have increasingly taken into account extra information to improve the recommendation quality, such as the location of uses if it is available. [29] presents a comprehensive overview of the recent progress in recommendation services in location-based social networks. The basic idea is that location helps to bridge the gap between the physical world and online social services. Research effort in this field can be categorized into four main groups based on the object to be recommended, including friends, locations, location-based activities, and location-based online services. The analyses utilize a variety of data sources such as user profiles, location histories (user trajectories), and geo-tagged social media activities, and include methods such as collaborative filtering, content-based recommendations, and network analysis. As examples, the works in [62] and [203] have shown that adding geographical information, e.g., imposing a gravity model on the probability of visiting different locations, can significantly improve the performance of recommending points-of-interest (POIs). These works demonstrate that using information about physical spaces helps in learning people's behavioral preferences. Consequently, there has been a number of recent studies in the literature that integrate data from different sources, such as location and other complementary information, in learning individual behaviors in the specific domain of product or app adoption [142, 205, 159, 108]. It is also worth mentioning that they have built different representations to capture the relationships between locations, POIs, and apps.

4.2.2 Geometric deep learning

Recent development in deep learning techniques [129] has mostly advanced the state-of-the-art in a variety of machine learning tasks. Classical deep learning approaches are most successful on data with an underlying Euclidean or grid-like structure with a built-in notion of invariance. Real-world data, however, often come with a non-Euclidean structure such as graphs and manifolds. However, it is not straightforward

to generalize them to cope with data that come with a non-Euclidean structure such as graphs and manifolds, mostly due to a lack of the notion of shift-invariance and well-defined operations such as convolution. To cope with these challenges, geometric deep learning [52] is a branch of emerging deep learning techniques that make use of novel concepts and ideas brought about by graph signal processing [176], a fast-growing field by itself, to generalize classical deep learning approaches to data lying in non-Euclidean domains such as graphs and manifolds.

Notable examples of recent development in geometric deep learning include [53, 70, 118], where the authors have defined the convolution operation on graphs indirectly via the graph spectral domain by making use of a generalized notion of spectral filtering, as well as the work of [152], where the authors propose a spatial-domain convolution on graphs using local patch operators represented as Gaussian mixture models. Out of the many successful applications, geometric deep learning techniques have been applied to the problem of matrix factorization with state-of-the-art performances in recommendation tasks [153, 190]. The present chapter was inspired by this line of research; however, two notable differences are that, we make use of external auxiliary and network information, and we utilize the graph attention networks to enforce meaningful local smoothness constraints on the solutions.

Attention-based models are inspired by human perception [148]: instead of processing everything at once, we process the task-relevant information by selectively focusing our attention. Since their inception, attention-based models have been successfully applied to several different deep learning tasks. Attention has been shown to help identify the most task-relevant parts of an input, ignore the noise, and interpret results [191]. Recently, attention mechanisms have been generalized for graph structured data [192], which opens possibility for designing various graph attention networks [132]. These graph-based attention models also enable the combination of data from multiple views [173]: in addition to improving model performance with more task-relevant data, using data from multiple views improves model interpretability by allowing us to learn which views are the most task-relevant from the learned attention weights.

4.3 Data description

We utilize the open data set provided by Yelp[2], which is an online review platform where users may rate and post reviews on businesses such as restaurants, bars, spas, etc. As a proof of concept, we focus the analysis in this section of my thesis on the cities of Cleveland Heights and Urbana in the United States, with 3989/4067 users and 196/392 businesses in the data set. The density of nonzeros in the matrix is 1.05% / 0.45%. We summarize the statistics of the ratings for the businesses in Table 4.1. We show the distributions of the number of reviews of businesses and users in Figure 4-3a and Figure 4-3c, respectively, both of which follow a power-law distribution. The distributions of the average ratings of businesses and users are shown in Figure 4-3b and Figure 4-3d, respectively, the former of which follows a unimodal distribution and the latter a multimodal distribution.

Table 4.1: **Summary statistics of the data.**

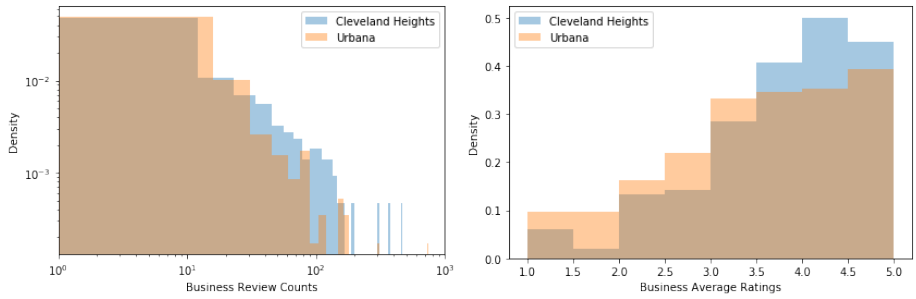
	Cleveland Heights	**Urbana**
Review count	6646	7255
User count	3989	4067
Business count	196	392
Average rating (Std.)	3.823 (1.310)	3.696 (1.439)
Reviews per user (Std.)	1.666 (2.195)	1.784 (2.405)
Reviews per business (Std.)	33.908 (57.584)	18.508 (46.395)

4.3.1 Business

Yelp collected information about businesses via both updates from business owners and surveys from users. The rich information of different nature makes it an ideal data set for testing the effectiveness of the proposed method. Roughly speaking, there are three types of information about the businesses, i.e., basic information (attributes and categories), location information, and check-in information (temporal popularity).

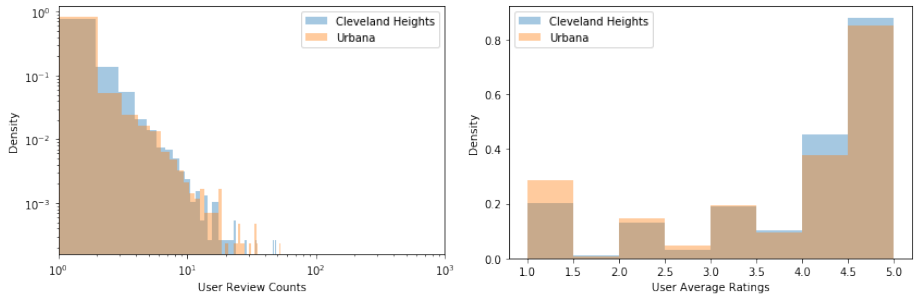
1. The basic information (business attributes and business categories). This includes star ratings (rounded to half-stars), review counts, business categories, and other

[2]The data set can be obtained through this link.



(a) business review counts

(b) business average ratings



(c) user review counts

(d) user average ratings

Figure 4-3: Distributions of average ratings of businesses and users.

attributes related to amenities. 77 business attributes cover information such as the provision of parking space, WiFi hotspot, and takeaway service. Since most attributes are categorical variables, we adopt one-hot (i.e., one-of-K) encoding that indicates whether the business belongs to the particular attribute or not. There are 369 business categories, e.g., Mexican, Burgers, Gastropubs, and each business may belong to multiple categories. Similarly, we adopt one-hot encoding on these categories.

2. The location information. The location information, in latitude and longitude, allows us to locate the businesses on the map as shown in Figure 4-4, where the color code represents the average rating of each business. There seems no obvious relationship between spatial proximity and similarity in average rating. To see this more clearly, we plot the relationship between distance and difference in average rating of businesses in Figure 4-5a. The box plots for differences in average rating for businesses of various distances are almost the same, except

for the case of very large distance. There also seems to be no relationship between spatial distance and cosine distance between the business attribute vectors mentioned above, as can be seen in Figure 4-5b.

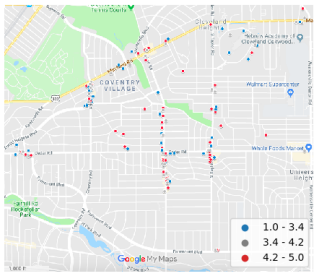
3. The check-in information from users. This allows us to analyze the temporal patterns of the popularity of the businesses. We aggregate the check-ins into 144 hourly bins of a week (24 h $\times$ 7 days) to obtain a check-in vector for each business. We then analyze the relationship between cosine distance between the check-in time vectors and 1) the difference in average rating (Figure 4-5c) as well as 2) the cosine distance between the business attribute vectors (Figure 4-5d). For both plots, when the cosine distance between the check-in time vectors is larger than 0.3, there is a slight increase in the y-axis in both plots. This seems to suggest that, when the difference in temporal popularity is larger, the difference in average rating and business attributes tends to be larger as well. One explanation for this might be that businesses with similar temporal popularity patterns attract customers with similar preferences.

In summary, the auxiliary information about each business is represented by a vector of dimension 592, which includes 446 features from the attribute and category information, 2 features from location, and 144 features from the check-in information.

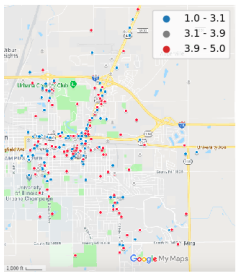
4.3.2 Users

Similarly, the information about users can be categorized into two types, i.e., basic metadata and friendships on Yelp.

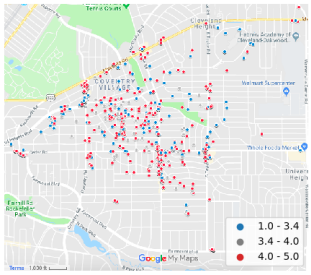
1. Basic metadata about each user. These include the number of "useful", "funny", and "cool" reviews, number of fans, number of compliments on reviews as being "hot", "cute", "plain", "cool", "funny", or "good writer", and number of compliments on the user's profile, lists, notes, photos, and other information. These lead to a total of 15 attributes (auxiliary information) for each user.



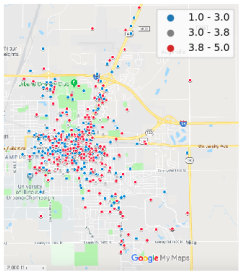
(a) Cleveland Heights: business



(b) Urbana: business



(c) Cleveland Heights: user

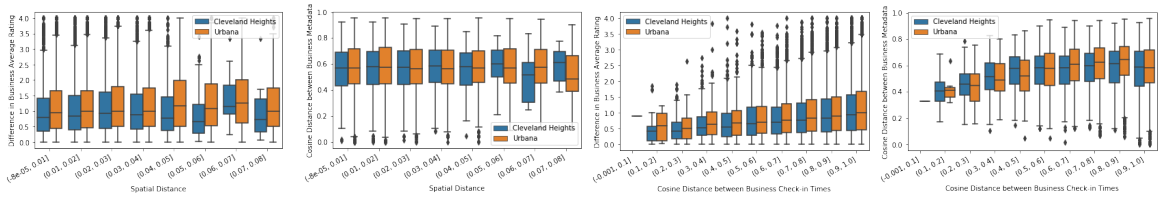


(d) Urbana: user

Figure 4-4: **Spatial distributions of businesses and users.** The spatial location of a user is the weighted average location of the businesses she has reviewed. The color code represents the average ratings of businesses and users.

2. The location information. We average the businesses users visited and locate them on the map as shown in Figure 4-4c and 4-4d for Cleveland Heights and Urbana respectively, where the color code represents the average rating each user scored for businesses. There seems to be no obvious relationship between spatial proximity and similarity in average rating.

3. Friend relationship. This provides a list of Yelp users as friends of each given user. With this information, we build a friendship network and link two users if they are friends with each other. We first analyze the count of user pairs according to different degrees of separation in the friendship network, as shown in Figure 4-6a. We see that the number of pairs increases from one to three degrees of separation and then decreases afterward. We further analyze the relationships between degrees of separations and 1) the difference in average user



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

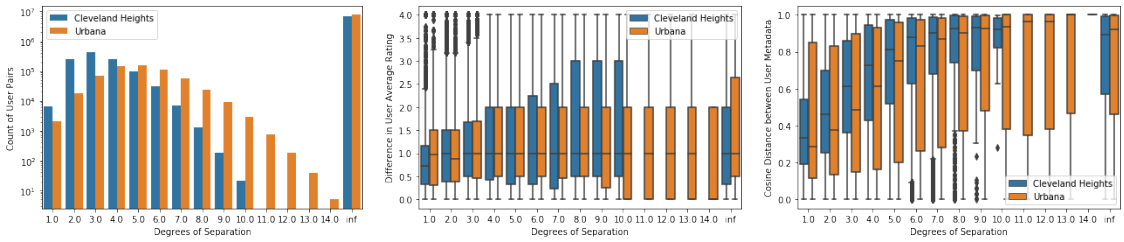
Figure 4-5: **Relationship between spatial distance, difference in average rating, cosine distance between the business attribute vector, and cosine distance between the check-in time vector.**
The middle line, lower and upper boundaries of the box (interquartile range or IQR) correspond to mean, median, 25% and 75% of the data, respectively. The lower and upper whiskers extend maximally 1.5 times of IQR from 25 percentile downwards and 75 percentile upwards, respectively.

rating (Figure 4-6b) as well as 2) the cosine distance between user metadata vectors (Figure 4-6c). There seems to be no obvious relationship between the two in our datasets. However, several studies have shown that social connections are useful predictors of one's underlying preferences [180, 144], and we suspect this that for different datasets, the closer two individuals in the Yelp friendship network, the more similar their average rating as well as metadata would be.

The preliminary analyses presented in this section provide an initial understanding of the potential usefulness of the diverse business and user information available on Yelp or similar data platforms. In practice, depending on the results of these analyses and the task at hand, we may consider a subset of the information available as input to our model, as explained in Section 4.5.

4.4 Model

In this section, we first present and formulate our problem. We then describe how the latent representations for user and business are learned and how to emphasize the relevant information for rating prediction. Finally, we present the proposed learning framework and algorithm.



(a) (b) (c)

Figure 4-6: **The count of user pairs, difference in average rating, and cosine distance between the user metadata vector, concerning the degrees of separation in the Yelp friendship network.** The middle line, lower and upper boundaries of the box (interquartile range or IQR) correspond to mean, median, 25% and 75% of the data, respectively. The lower and upper whiskers extend maximally 1.5 times of IQR from 25 percentile downwards and 75 percentile upwards, respectively.

4.4.1 Problem formulation

We now formally define our inference problem. We consider a partially observed user-business rating matrix $X \in \mathbb{R}^{n \times m}$, where n is the number of individuals, m is the number of businesses, and the ij-th entry X_{ij} is the rating of user i on business j, where $X_{ij} \in \{1, 2, 3, 4, 5\}$. We consider auxiliary information about users and businesses, as discussed in section 4.3.2 and 4.3.1, which are denoted as $X_{(U)} (\in \mathbb{R}^{4797 \times 15})$ and $X_{(B)} (\in \mathbb{R}^{929 \times 592})$, respectively. We further build a user-user and a business-business network to capture the relationships among users and businesses. The user network is defined as the Yelp friendship network, whose adjacency matrix is denotes as G_U, and the corresponding combinatorial graph Laplacian matrix as L_U. For businesses, due to lack of external relational information, we build a 10-nearest neighbor similarity network G_B, in which a binary edge between businesses i and j indicates that i is among the 10 businesses that are closest to j (or vice versa) in terms of the cosine similarity between the corresponding row vectors of $X_{(B)}$ and $X_{(B)}$. The notations used in this study are summarized in Table 4.2.

Our objective is to infer the users' preferences on the businesses they have not yet rated, i.e., to complete the empty entries in X given the observed entries as well as the complementary information provided by G_U, G_B, $X_{(U)}$, and $X_{(B)}$. We cast this

problem as a matrix completion problem given additional relational (network) and non-relational information. One of the variants of the matrix completion problem is to find a low-rank matrix $\hat{X}$ that matches the original matrix X conditioned on the observed entries. In practice, for robustness against noise as well as computational efficiency, the problem is often formulated using matrix factorization [177]. The corresponding objective function is:

$$\min_{U,V} ||\mathbf{\Omega}_{\text{obs}} \circ (X - UV^T)||_F^2 + \mathcal{R}(U) + \mathcal{R}(V), \tag{4.1}$$

where $U \in \mathbb{R}^{n \times k}$ and $V \in \mathbb{R}^{m \times k}$ are two latent space representations, with k being the dimension of the latent space, $\mathbf{\Omega}_{\text{obs}}$ is an indicator matrix with 1 at the observed entries of X and 0 otherwise, $\circ$ denotes the Hadamard product, and $|| \cdot ||_F$ denotes the Frobenius norm. The regularization terms $\mathcal{R}(U)$ and $\mathcal{R}(V)$ enforce additional constraints on the structure of U and V. In our context, we interpret U as a latent representation that captures the preferences of users on businesses, and V as a latent representation that encodes the characteristics of businesses. In this section of my thesis, we focus on solving the problem of Eq. (4.1) where the choices of $\mathcal{R}(U)$ and $\mathcal{R}(V)$ enforce local and global smoothness of U and V. We discuss these choices in more detail in the following section.

4.4.2 Local smoothness regularization with graph attention network

In the problem of Eq. (4.1), it is common to consider a regularization term $\mathcal{R}(\cdot)$ that enforces certain structure of the solution. The common forms of $\mathcal{R}(\cdot)$ include L_2 norm, L_1 norm, and smoothness with respect to some underlying network structure. In our context, for example, the smoothness of the latent user representation U can be promoted by adding a regularization term in the form of $\mathcal{R}(U) = \text{tr}(U^T L_U U)$, where L_U is the graph Laplacian matrix of a user friendship graph, and $tr(\cdot)$ denotes the trace operator. Such a regularization term in Eq. (4.1) enforces that the representations for users who are neighbors in the network are close to one another in the latent space.

Table 4.2: **Notations**

Notations.	**Definitions and Descriptions**
X	User-business rating matrix
$\hat{X}$	Predicted user-business rating matrix
U	Inferred latent user representation
V	Inferred latent business representation
$X_{(U)}$	Auxiliary information about users
$X_{(B)}$	Auxiliary information about business
G_U	User friendship network
L_U	Graph Laplacian of the user friendship network
G_B	Business similarity network
L_B	Graph Laplacian of the Business similarity network
$a(\cdot)$	Attention mechanism
$\mathbf{h}$	Hidden units
$f_l(\cdot)$	Linear layer in the neural network
LeakyReLU$(\cdot)$	Leaky Rectified Linear Unit activation function
ELU$(\cdot)$	Exponential Linear Unit activation function
softmax$(\cdot)$	Softmax activation function
T	Number of iterations in the training
T_R	Number of temporal steps in the LSTM layer

The smoothness constraint described above is a global one in the sense that it enforces the representations for every pair of friends to be similar across the entire network. While this is a reasonable and widely adopted assumption in the context of recommendation, in practical situations the edges in the observed friendship network are not necessarily all meaningful or of equal importance. In this case, a local smoothness constraint may be more appropriate, i.e., only a subset of friends would affect a given user's preference. In this section of my thesis, we propose to promote such local smoothness of the solutions U and V using the graph attention network (GAT) [192]. As illustrated in Figure 4-7, the GAT places higher weights on neighbors who provide task-relevant information and lower weights on those who do not. In other words, neighbors are not weighted equally but instead by how they contribute to the recommendation (matrix completion) task. This leads to two key benefits of the proposed framework: 1) removing noisy connections as well as weighing relevant neighbors differently in the network; 2) revealing how information is aggregated via the attention weights, hence rendering the framework more interpretable.

We now explain the GAT in more detail. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V}$ as the node set and $\mathcal{E}$ the edge set, we first define the concepts "node embedding" and "edge embedding" as follows.

Definition 4.4.1. ***Node embedding*** *[56]. A node embedding is a function $g_n : v \to \mathbb{R}^k$, which maps each node $v \in \mathcal{V}$ to a k-dimensional vector where $k \ll |\mathcal{V}|$.*

Definition 4.4.2. ***Edge embedding*** *[56]. An edge embedding is a function $g_e : e \to \mathbb{R}^{k'}$, which maps each edge $e \in \mathcal{E}$ to a k'-dimensional vector.*

We now describe the graph attention mechanism for enforcing local smoothness constraints in our framework. Without loss of generality, we focus on the update of the latent user representation U, but the same procedure applies directly to the update of the latent business representation V. A single GAT layer in a neural network architecture takes as input a user network G_U in which nodes represent users, and a node embedding matrix denoted as $U^T = [\mathbf{h}_1^T, \mathbf{h}_2^T, ..., \mathbf{h}_n^T]$, $\mathbf{h}_i^T \in \mathbb{R}^{d_i}$, where n represent the number of nodes (users) in the layer. The output is a new node embedding matrix updated via the attention mechanism. More specifically, from the perspective of v_i, the attention mechanism first takes as input the current node embedding for v_i and its neighbors in G_U, and compute an edge embedding for each edge between v_i and a neighbor v_j. This step is explained as follows:

$$v \to e : \alpha_{ij} = \text{softmax}_j(a(f_l(\mathbf{h}_i), f_l(\mathbf{h}_j))) = \text{softmax}_j\Big(\text{LeakyReLU}\big(\big[f_l(\mathbf{h}_i)||f_l(\mathbf{h}_j)\big]\mathbf{a}\big)\Big), \tag{4.2}$$

where $||$ represents concatenation. In Eq (4.2), α_{ij} is computed via a shared (self-)attention mechanism $a(\cdot) : \mathbb{R}^{d_o} \times \mathbb{R}^{d_o} \to \mathbb{R}$, where d_o is the dimension of the linearly transformed input node embedding $f_l(\mathbf{h}_i)$, and $\mathbf{a} \in \mathbb{R}^{2d_o}$ is a coefficient vector. We then normalize α_{ij} across all neighbors v_j of v_i using a softmax function as in Eq (4.2). This step can therefore be regarded as a mapping from node embeddings U to edge embeddings $\{\alpha_{ij}\}$, as illustrated in Figure 4-7, where α_{ij} is the attention weight that determine how much attention should be paid to a particular neighbor v_j when updating information on v_i. The second step in the attention mechanism is then

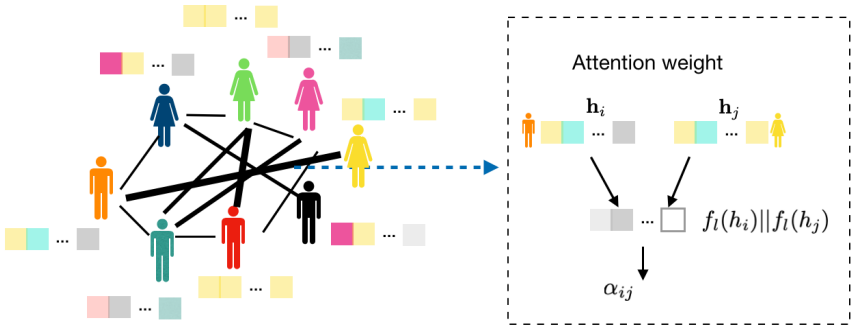


Figure 4-7: **Basic idea behind a graph attention network.** The attention weight α_{ij} represents the relevance of neighbor v_j in updating information on v_i.

a mapping from node embedding to node embedding using the obtained attention weights $\{\alpha_{ij}\}$:

$$e \to v : \mathbf{h}_i^{\text{new}} = \text{ELU}\Big(\sum_j f_l\big(\alpha_{ij}\mathbf{h}_i\big)\Big). \tag{4.3}$$

The attention mechanism described in Eq (4.2) and Eq (4.3) is referred to as a single-head attention. To stabilize the learning process, we follow the ideas in [191, 192] to perform a multi-head attention to update the node embedding:

$$\mathbf{h}_i^{\text{new}} = \text{ELU}\Big(\frac{1}{K}\sum_{k=1}^{K}\sum_j f_l\big(\alpha_{ij}^{(k)}\mathbf{h}_i\big)\Big), \tag{4.4}$$

where $\{\alpha_{ij}^{(k)}\}$ are the attention weights from the k-th independent attention mechanism of Eq (4.2).

4.4.3 Framework

We are now ready to present the proposed framework for the matrix completion task. Specifically, we propose to solve the following optimization problem:

$$\begin{aligned} &\min_{U,V} \ ||\mathbf{\Omega}_{\text{obs}} \circ (X - UV^T)||_F^2 + \alpha \ \text{tr}(U^T L_U U) + \beta \ \text{tr}(V^T L_B V), \\ &\text{s.t. } U^{(t)} \longleftarrow \text{GAT}(G_U, U_{T^R}^{(t)}), \ V^{(t)} \longleftarrow \text{GAT}(G_B, V_{T^R}^{(t)}), \end{aligned} \tag{4.5}$$

where $U^{(t)}$ and $V^{(t)}$ are the updates of U and V after iteration t of the proposed learning architecture (see Figure 4-8 and Algorithm 1), $U_{T^R}^{(t)}$ and $V_{T^R}^{(t)}$ are the intermediate output of the long-short-term-memory (LSTM) layer after iteration t, and α and β are two hyperparameters. The objective in Eq. (4.5) follows that in Eq. (4.1), where $\mathcal{R} = \text{tr}(U^T L_U U)$ and $\mathcal{R} = \text{tr}(V^T L_B V)$ are two global smoothness constraints imposed on the solutions U and V, respectively. More importantly, we constrain that the updates of U and V after each iteration are both obtained via a GAT layer, which further adds local smoothness properties to the solutions. On the one hand, the global smoothness constraints promote the behavior of the algorithm that similar users or businesses are mapped to close-by positions in the latent spaces, i.e., ratings of friends or similar businesses are generally similar. On the other hand, the local smoothness constraints allow the algorithm to update the latent representations by discarding noisy connections in the observed user and business networks and selectively aggregating information from neighbors according to their relevance.

To solve the problem of Eq. (4.5), we build upon [153] and propose a novel geometric deep learning architecture, as illustrated in Figure 4-8. We name the proposed framework as Multi-Graph Graph Attention Network (MG-GAT). The MG-GAT architecture consists of three layers: the dense layer, the LSTM layer, and the GAT layer. We first define the input to the dense layers in the MG-GAT. First, U_{svd} and V_{svd} are obtained by applying singular value decomposition (SVD) to the observed matrix X. We then define U_{concat} and V_{concat} as $U_{\text{svd}}||X_{(U)}$ and $V_{\text{svd}}||X_{(B)}$, respectively. The dense layer takes as input U_{concat} for the user case, and V_{concat} for the business case, and maps them to a new feature space. The next layer is the LSTM layer, as suggested by [153], which allows small changes in the latest updates of U and V to pass through the temporal steps within the LSTM. The inputs to the LSTM layer are the output of the previous dense layer and that of the previous GAT layer. The third layer is the GAT layer, which takes as input the output of the LSTM layer as well as the corresponding network information, and updates U and V by selectively aggregating information from neighbors according to their relevance. This process is repeated for both U and V and, upon the final iteration, the predicted (completed)

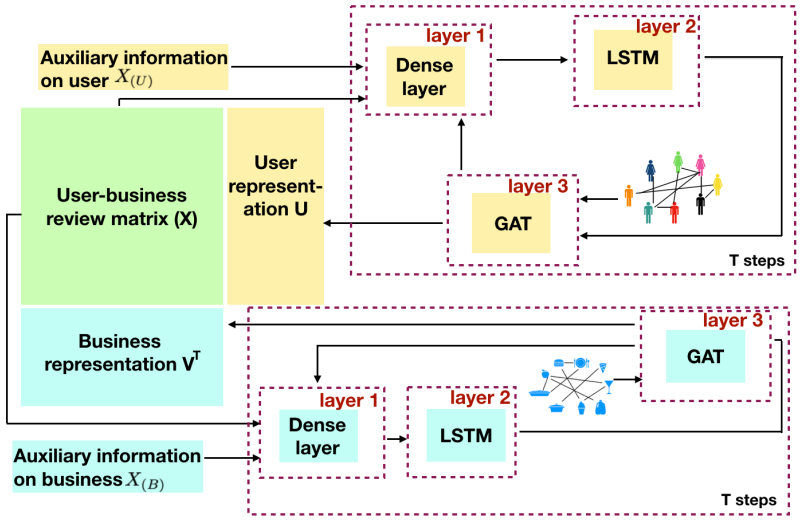


Figure 4-8: **The proposed geometric deep learning architecture for learning latent user representations U and business representations V, via iterations over three layers: a dense layer, an LTSM layer, and a GAT layer.** The predicted $\hat{X}$ is obtained using the final updates of U and V via $\hat{X} = UV^T$.

matrix $\hat{X}$ is obtained via $\hat{X} = UV^T$. We use Adam stochastic optimization to train the model and learn the parameters [117], and the complete algorithm is presented in Algorithm 1.

4.5 Results

4.5.1 Experimental setting

We split the rating data into training, validation, and test sets according to temporal information about the ratings, i.e., ratings between 2009 and 2016 as the training set, ratings between 2016 and 2017 as the validation set, and ratings between 2017 and 2018 as the test set. We use the average Root Mean Squared Error (RMSE) as the performance metric:

$$\text{RMSE} = \sqrt{\frac{||\mathbf{\Omega}_{\text{tes}} \circ (X - \hat{X})||_F^2}{|\mathbf{\Omega}_{\text{tes}}|_1}} = \sqrt{\frac{||\mathbf{\Omega}_{\text{tes}} \circ (X - UV^T)||_F^2}{|\mathbf{\Omega}_{\text{tes}}|_1}}, \tag{4.6}$$

Algorithm 1: Multi-Graph Graph Attention Network (MG-GAT)

input X, $X_{(U)}$, $X_{(B)}$, G_U, G_B, T, T^R, k;
Initialization $U^{(0)}$, $V^{(0)}$ (initialized with uniform distribution)
$(U_{\text{svd}} \in \mathbb{R}^{n\times k}, V_{\text{svd}} \in \mathbb{R}^{m\times k}) = \text{svd}(X)$
$U_{\text{concat}} = U_{\text{svd}}||X_{(U)}$, $V_{\text{concat}} = V_{\text{svd}}||X_{(B)}$
for $t = 1 : T$ **do**
    Update the user representations
    Feed U_{concat} into the dense layer and produce an output $\tilde{U}^{(t)}_{\text{concat}}$
    Set $U^{(t)}_0 = U^{(t-1)}$.
    for $j = 1 : T^R$ **do**
        Feed $U^{(t)}_{j-1}$, $\tilde{U}^{(t)}_{\text{concat}}$ into LSTM and produce $U^{(t)}_j$.
    end
    Feed $U^{(t)}_{T^R}$ and G_U to GAT to produce an output $U^{(t)} \in \mathbb{R}^{n\times k}$
    Update the business representations
    Feed V_{concat} into the dense layer and produce an output $\tilde{V}^{(t)}_{\text{concat}}$
    Set $V^{(t)}_0 = V^{(t-1)}$.
    for $j = 1 : T^R$ **do**
        Feed $V^{(t)}_{j-1}$, $\tilde{V}^{(t)}_{\text{concat}}$ into LSTM and produce $V^{(t)}_j$.
    end
    Feed $V^{(t)}_{T^R}$ and G_B to GAT to produce an output $V^{(t)} \in \mathbb{R}^{m\times k}$
output $U \leftarrow U^{(T)}, V \leftarrow V^{(T)}, \hat{X} = UV^T$

where X is the original user-business rating matrix, $\hat{X}$ is the predicted matrix, $\mathbf{\Omega}_{\text{tes}}$ is the indicator matrix with 1 for the entries in the test set and 0 otherwise, and $|\cdot|_1$ represents entry-wise L_1 norm.

The auxiliary information presented in Section 4.3 about users and businesses may be selectively considered as input to Algorithm 1. In this study, we choose to only use auxiliary information about businesses, i.e., $X_{(B)}$, which led to better performance.

We now describe how we determine and tune the model parameters. In our experiments, we set $T = 5000$ and $T^R = 10$. We tuned the rank over the values 4, 8, 16, 32, and 64. The learning rate controls how fast the neural network updates the weights, and is tuned using random search within the range of $[10^{-6}, 10^0]$. The dropout rate (i.e., the probability that a random neuron is ignored during training) controls the amount of regularization and is tuned within the range of $[0, 1]$. The hyperparameters α and β in the objective function of Eq. (4.5) are tuned within the

range of $[10^{-16}, 10^{0}]$. The best set of parameters we found for Cleveland Heights were a rank of 8, a learning rate of 0.011512, a dropout rate of 0.027535, $\alpha = 2.109724e-08$, and $\beta = 2.254494e-08$. The best set of parameters we found for Urbana were a rank of 64, a learning rate of 0.003009, a dropout rate of 0.648956, $\alpha = 3.647857e-09$, and $\beta = 7.119423e-09$.

We consider the following baseline models in the performance comparison. We first consider classical approaches including *Singular Value Decomposition (SVD)* [7], *non-negative matrix factorization (NMF)* [130], and *principle component analysis (PCA)* [100], which are the building blocks for many recommender systems. We then consider a recent approach, i.e., Multi-Graph Convolutional Neural Network (MGCNN), which is a state-of-the-art method for matrix completion proposed in [153]. Similarly to the proposed MG-GAT algorithm, these benchmarks extract latent representations for users and businesses given the observed data matrix.

4.5.2 Learning performance

The methods we test can be broadly grouped into non-deep learning based methods (SVD, PCA, and NMF) and deep learning based methods (MGCNN, and MG-GAT with its variants). The different input information required and methodologies adopted in these approaches allow us to evaluate their impact on the learning performance.

Table 4.3: **Performance comparison using RMSE metric (standard errors in parentheses).**

Category	Method	Cleveland Heights	Urbana
Non-deep learning	NMF	2.341 (1.922e-13)	2.665 (7.221e-17)
	PCA	1.366 (0.001758)	1.536 (1.630e-05)
	SVD	1.362 (1.163e-08)	1.542 (5.741e-09)
Deep learning benchmark	MGCNN [153]	1.367 (0.002332)	1.528 (0.001652)
MG-GAT	MG-GAT	**1.330 (0.003324)**	**1.397 (0.002501)**
	MG-GAT (shuffled edges)	1.331 (0.001814)	1.428 (0.003304)
	MG-GAT (missing edges)	1.347 (0.003489)	1.416 (0.003521)

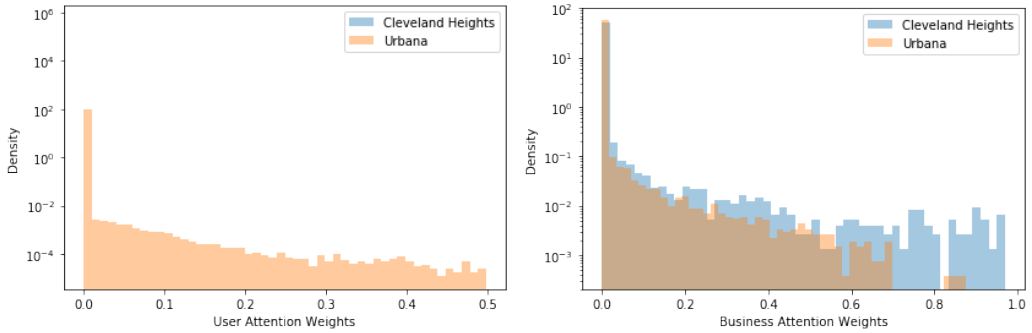
$MG-GAT(no$

We summarize the performance of our method and the baselines in Table A.1. First,

we observe that deep-learning based methods perform better than non-deep learning based methods, which highlights the benefit of data-driven learning. Second, the proposed method MG-GAT outperforms MGCNN, which is due to 1) the incorporation of auxiliary information about businesses, which is not taken into account in MGCNN; and 2) the introduction of the attention mechanism that is able to select the most relevant information in learning the latent representations.

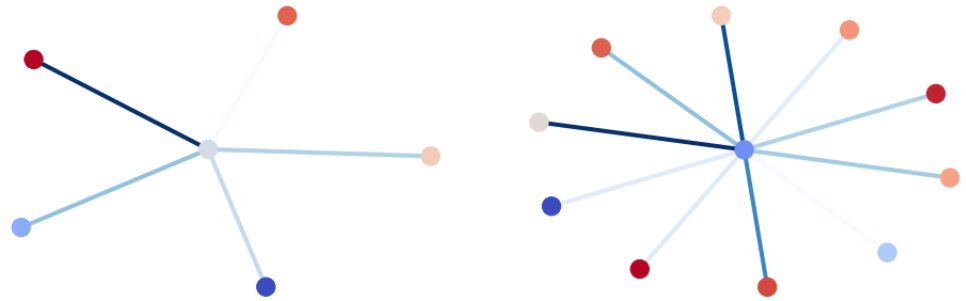
We further compare three variants of the proposed method with different input information. If we drop all the auxiliary information, we see that the learning performance is the worst among all variants. This again indicates the importance of integrating heterogeneous information sources in prediction, in this case the information about businesses described in section 4.3.1. We then consider another two variants where we (1) randomly drop 50% of the edges in the observed user and businesses networks, or (2) randomly shuffle 50% edges in both networks. We see that both missing edges and shuffled edges negatively affect the learning performance. When the information contained in the network is highly informative, such as in the case of Urbana, there exists a larger gap in the performance between the case of complete edges and that of missing or shuffled edges. Moreover, the networks with shuffled edges, which provide noisy information in this case, perform worse than that with missing edges. For Cleveland Heights, we see that networks with shuffled edges actually outperform that with missing edges, which indicates that in this case edges (and in particular user relationships given the larger number of users compared to businesses) are not particularly informative. As we see below, this is reflected in the weights learned by the graph attention networks, which largely ignore such noisy information.

Finally, we analyze the attention weights used for the final update of the latent user and business representations. We show the distribution of attention weights for both cases in Figure 4-9. For Urbana (shown in orange), while most of the attention weights are zero (due to absence of edges between node pairs), there is a decay of frequency of edges as attention weights increase. For Cleveland Heights, all user attention weights are negligibly small , which is in line with the results described above suggesting that the social network does not provide much information in this particular case. In



(a) attention weights for user representations(b) attention weights for business representations

Figure 4-9: **Distribution of attention weights.**



(a) (b)

Figure 4-10: **Attention weights for a focal (a) user and (b) business.** Nodes are colored by average rating of the user or business, and the intensity of a link represents the attention weight.

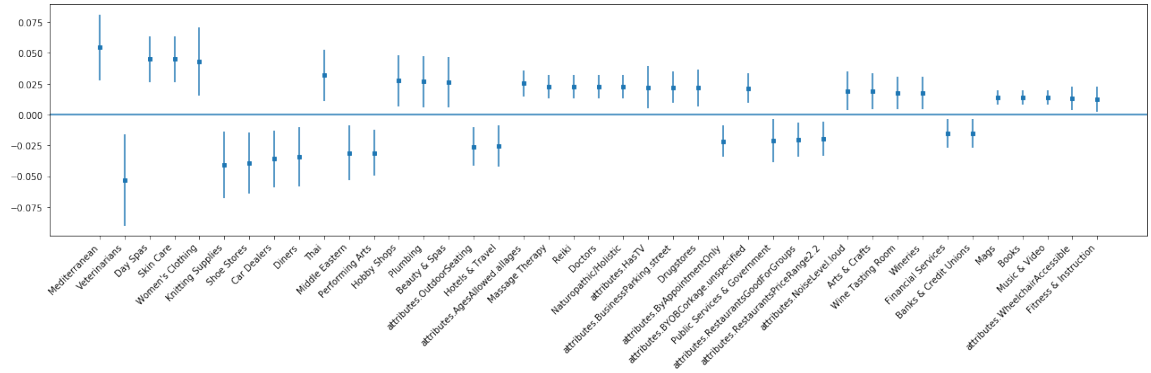
general, the distributions of weights shown in Figure 4-9 suggest that the attention mechanism is able to identify relevant neighbors for each node and assign weights accordingly, which is clearly different from the setting without attention, where the weights on all the edges are effectively the same. We also illustrate the attention weights for a focal user (Figure 4-10a) and a focal business (Figure 4-10b), in which we observe that the neighbors are weighted differently. This contributes to the superior performance of the proposed method.

4.5.3 Pattern analysis on business representations

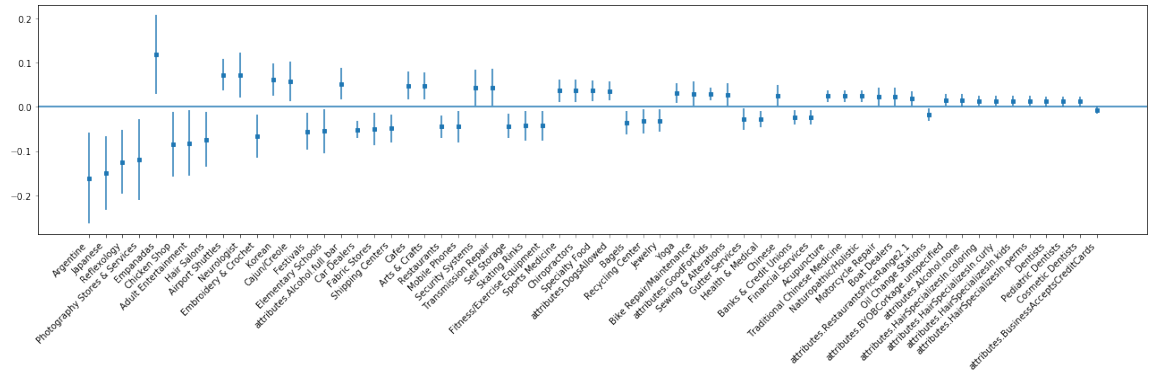
Information contained in latent business representations

We are first interested in understanding the information contained in the learned latent business representations V, e.g., with respect to business attributes and categories. To this end, we apply eigendecomposition to the covariance matrix V^TV, and observe that the leading eigenvector explains most of the variances in V, indicating that the information contained in the business representations is mainly concentrated in one dimension. In this case, we found that the Pearson correlation coefficient between the leading eigenvector and the vector that contains the average rating of each business is 0.549 (with p-value of 7.742e-17) for Cleveland Heights and 0.818 (with p-value of 1.325e-95) for Urbana.

To better understand the information contained in this leading eigenvector, we apply a regression analysis where the dependent variable is the eigenvector, and the independent variables are binary variables that indicate presence of certain business attributes and categories. In Figure 4-11, we visualize the coefficients and the corresponding confidence intervals sorted decreasingly by the magnitude of the coefficients. Interestingly, the importance of businesses categories in explaining the leading eigenvector differs. For Urbana, categories of high importance are mostly related to restaurant, while for Cleveland Heights, they are mainly related to leisure. For Urbana, *empanadas*, *airport shuttles*, *neurologist*, *cajun* and *Korean restaurant* contribute positively to the leading eigenvector, while *Argentine*, *Japanese*, *reflexology*, *photography stores & services*, *embrodiery* and *crochet* contribute negatively. For Cleveland Heights, *Mediterranean*, *day spas*, *skincare*, and *women's clothing* contribute positively to the leading eigenvector, while *veterinarians*, *knitting supplies*, *shoe stores*, *Car rentals*, and *dinners* contribute negatively. These results demonstrate that the learned latent representation for businesses capture meaningful information that can be interpreted in terms of business categories or attributes.



(a) Cleveland Heights



(b) Urbana

Figure 4-11: **Regression coefficients with confidence intervals for explaining the leading eigenvector of V^TV.**

Clusters of businesses using latent business representations

Next, we apply cluster analysis to businesses using the learned latent business representations. More specifically, we perform k-means clustering on the learned representation V, where we use the elbow method [93] to identify the optimal number of clusters to be 3. We label the three clusters as low-rating, medium-rating, and high-rating clusters, according to the average rating of businesses in each cluster. The statistics of ratings for businesses in the three clusters are summarized in Table 4.4.

Table 4.4: **Cluster characteristics for Cleveland Heights (C.H.) and Urbana (U.).**

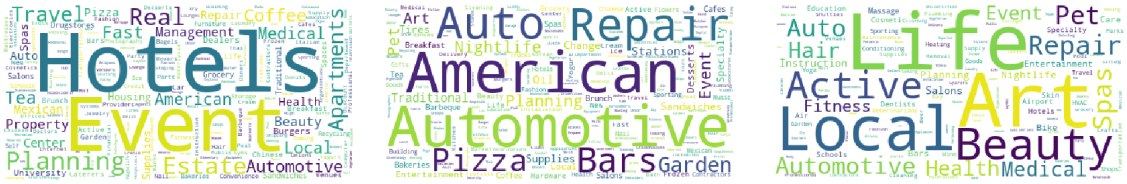
	Cluster 1 (C.H./U.) (low)	**Cluster 2 (C.H./U.) (medium)**	**Cluster 3 (C.H./U.) (high)**
count	14/138	118/140	64/114
mean	2.33/2.43	3.52/3.69	4.45/4.44
standard deviation	1.14/0.80	0.78/0.60	0.56/0.51
25 percentile	1.39/1.85	3.17/3.38	4.19/4.15
50 percentile	2.19/2.53	3.63/3.71	4.50/4.50
75 percentile	2.79/2.99	4.00/4.00	5.00/4.87

We then analyze the categories of businesses in each cluster and visualize the frequency of different business categories by the word clouds in Figure 4-12 and 4-13 for Cleveland Heights and Urbana, respectively. The larger the name of the category, the more frequent it appears in the cluster. Certain categories that predominantly appear in all three clusters in both cities, i.e., *service*, *restaurants*, *shopping*, *food*, *home*, and *store*, have been removed from the word clouds to better highlight the difference between the three clusters.

As we can see, the business categories in low, medium and high-rating clusters are different for Cleveland Heights and Urbana. For Cleveland Heights, the frequent categories in low-rating clusters, shown in Figure 4-12a, are *real estate, stations, drugstores, health*, and *supplies*. In medium-rating cluster shown in Figure 4-12b, the frequent categories include *nightlife*, *bar*, *American restaurant*, *pet*, and *traditional*. In the high-rating cluster in Figure 4-12c, we see that the frequent categories include *art*, *spas*, *beauty*, *entertainment hair*, and *salons*. For Urbana, the frequent categories in low-rating clusters, shown in Figure 4-13a, are *hotels* and event. In medium-rating cluster shown in Figure 4-13b, the frequent categories include *automotive*, *American*,

(a) Cluster 1 (low-rating) (b) Cluster 2 (medium-rating) (c) Cluster 3 (high-rating)

Figure 4-12: **Analysis of business categories of different business clusters in Cleveland Heights.** The relative sizes of the words correspond to the frequencies of the category.



(a) Cluster 1 (low-rating) (b) Cluster 2 (medium-rating) (c) Cluster 3 (high-rating)

Figure 4-13: **Analysis of business categories of different business clusters in Urbana.** The relative sizes of the words correspond to the frequencies of the category.

auto, *repair*, and *bars*. In the high-rating cluster in Figure 4-13c, we see that the frequent categories include *life*, *local*, *art*, *beauty* *active*, and *repair*. These results again demonstrate that the learned latent representation for businesses are related to business categories or attributes.

4.6 Conclusion and managerial implications

Recommender systems have significantly benefited and influenced both the consumers and businesses. They lie at the center of the decision-making of many online and offline consumer-based companies, such as Amazon, Netflix, and Yelp. The recent availability of large-scale consumer and business data present both opportunities and challenges in developing advanced data analysis frameworks to leverage multiple sources of heterogeneous information for more accurate, personalized and targeted recommendations.

In this section of my thesis, we propose a novel geometric deep learning framework for personalized recommendation by predicting users' preferences on businesses that they have not yet rated. Our framework possesses the following advantages over existing solutions. First, it is capable of handling multiple sources of heterogeneous information, including spatial and temporal information, relational (network) information, and other types of metadata on users or businesses. The consideration of relational information is aligned with the classical idea of exploiting user or business similarity; however, the contribution of the proposed framework lies in the encoding of such similarity in the form of networks that are directly incorporated into the analysis framework via machine learning models. This thus provides a general framework for analyzing data that come with topological structure. Second, the proposed method effectively filters information depending on its relevance, and assigns larger weights on information that is more predictive of the user preference. This unique feature is enabled by adding a local smoothness constraint, instead of a global one as normally considered in the literature (i.e., global similarity between users or businesses), imposed by the graph attention networks. This therefore provides a general way of selectively aggregating the most relevant and useful information for the task at hand. Testing our method on the rich Yelp data set, we demonstrate that both the direct incorporation of network structure and selective aggregation of information from relevant neighbors in the network are important to the learning performance in a prediction task. At the same time, the learned representations for businesses and users may be interpreted using domain knowledge. For example, we cluster the businesses using the business representations V and obtain three clusters of businesses of low, medium, and high ratings, which also seem to be associated with certain different business categories.

Our framework has several marketing applications and managerial implications. First, it can be applied in recommendation scenarios where companies would like to exploit the increasingly rich information at their disposal. Indeed, although the framework is presented in the context of Yelp, it can be applied to other online and offline platforms where multiple sources of (in particular relational) information are available about users and/or businesses. This enables a holistic understanding

and a more accurate prediction of consumer preferences, which are key to high quality recommendations. Second, our method is particularly useful when managerial decisions need to be made by looking at information that is specifically relevant to each user or business. For example, not all the friends of a given user will contribute equally to the understanding of that user's preferences. It is therefore important to filter out noise and only focus on the relevant information in learning meaningful patterns, and the proposed method can be regarded as an automatic pipeline to achieve this objective. Third, in addition to producing relevant recommendations based on prediction of user preferences, it would be important for companies to be able to interpret the learned patterns, so that business decisions can be made together with domain knowledge. The latent user and business representations enable such an interpretation, and the meaningful patterns extracted can be further used in other analysis tasks such as customer stratification, business grouping, link prediction, or even a refined understanding of user or business relationship via network inference [182]. The latter is particularly interesting as it provides a way of discovering the roles of users or businesses in the network [167], and the obtained network information can also be important to the design of efficient intervention or incentivization strategies [82].

There are several directions that are worth considering for future studies. First, there are scenarios where there exist multiple types of relationships between consumers or businesses, e.g., user relationship may be described in terms of their membership to different social groups. A method to integrate and weigh different networks will make the recommendation more powerful. Second, the objective of this study is to make recommendations that best predict users' preferences. However, some businesses may be further interested in maximizing revenue making use of the inferred preferences. The maximization of business revenue can thus be incorporated into the objective function of the learning framework. Lastly, the input information that our framework handles is currently static. One interesting direction is to develop adaptive and online learning framework that is able to handle and incorporate temporal data and dynamic relational (network) information.

Chapter 5

Learning Quadratic Games on Networks

Individuals, or organizations, cooperate with or compete against one another in a wide range of practical situations. In the economics literature, such strategic interactions are often modeled as games played on networks, where an individual's payoff depends not only on her action but also that of her neighbors. The current literature has largely focused on analyzing the characteristics of network games in the scenario where the structure of the network, which is represented by a graph, is known beforehand. It is often the case, however, that the actions of the players are readily observable while the underlying interaction network remains hidden. In this section of my thesis, we propose two novel frameworks for learning, from the observations on individual actions, network games with linear-quadratic payoffs, and in particular the structure of the interaction network. Our frameworks are based on the Nash equilibrium of such games and involve solving a joint optimization problem for the graph structure and the individual marginal benefits. Both synthetic and real-world experiments demonstrate the effectiveness of the proposed frameworks, which have theoretical as well as practical implications for understanding strategic interactions in a network environment.[1]

[1]This work is joint with Yehonatan Sella, Rodrigo Ruiz, and Alex Pentland.

5.1 Introduction

We live in an increasingly connected society. First studied by the American sociologist Stanley Milgram via his 1960s experiments and later popularized by John Guare's 1990 eponymous play, the theory of "six degrees of separation" has been recently re-analyzed on the social networking site Facebook, only to find out that any pair of Facebook users can actually be connected via approximately three and a half other ones [18]. Individuals, unsurprisingly, are not merely connected; their decisions and actions often influence the ones around them. Indeed, Christakis and Fowler [64] have found in a series of studies that, one's emotion, health habit, and political opinion can affect individuals who are as far as three degrees of separation in her social circle. Furthermore, such influence on the decision-making process may take place via either explicit [14, 181] or implicit interactions [22, 74].

To study the decision-making of a group of interacting agents, recent literature in economics has increasingly focused on the modeling of such interactions as games played on networks [110, 49]. The underlying assumption in this setting is that, in a game played by a group of players who form a social network, the payoff of a player depends on her action, e.g., an effort made to achieve a specific task, as well as that of her neighbors in the network. Two types of actions have been studied in the literature, i.e., strategic complements and strategic substitutes. In the former case, one's action increases her neighbors' incentives for action, e.g., students putting an effort together into a joint assignment or firms working on a collaborative research project [94]. In the latter case, however, the situation is opposite, such as the scenarios of firms competing on market prices or individuals on local public goods [48].

In a network game, the underlying structure of the network carries critical information and dictates the behavior and actions of the players. Typically, graphs are used as mathematical tools to represent the structure of these networks, and the current literature in this area has predominately focused on studying the characteristics of games on known or predefined graphs [21, 50, 82]. However, it is increasingly common that while ample observations on the actions of the agents are available, the underlying

complex relationships among them, which may be captured by an interaction network, remains mostly hidden due to cost in observation or privacy concern. In this case, the network needs to be estimated to better understand the present and predict the future actions of these agents. The primary goal of this section of my thesis is therefore to study the problem of learning, given the observations on the actions of the agents, a graph structure that best explains the observed actions in the setting of a network game.

Such a problem, generally speaking, may be thought of as an instance of the ones of learning relationships, often in the form of graph structures, from observations made on a set of data entities. Classical approaches from the machine learning and signal processing communities tackle this problem by building statistical models (e.g., probabilistic graphical models [121, 81]), physically-motivated models (e.g., diffusion processes on networks [90, 89]), or more recently signal processing models [73, 146]. These approaches, however, do not take into account the game-theoretic aspect of the decision-making of players in a network environment.

In the computer science literature, network games are known as graphical games [115] and there has been a few studies recently on learning the games from observed action data. For example, the works in [107, 104, 86, 87] have proposed to learn graphical games by observing actions from linear influence games with linear influence functions, where [88] has considered polymatrix games with pairwise matrix payoff functions. The work in [84] has proposed to learn potential games on tree-structured networks of symmetric interactions. These conditions have been relaxed in [85] where the authors have studied aggregative games where a player's payoff is convex and Lipschitz in an aggregate of their neighbors' actions defined via a local aggregator function. All these works, however, either consider a binary or a finite discrete action space, which may be restrictive in certain practical scenarios where actions take continuous values. Very recently, [30] has considered learning continuous-action graphical games, which is similar in spirit to our study albeit under a slightly different action (which is budgeted) and payoff setting.

In this study, we focus on learning games with linear-quadratic payoffs [21, 50, 5, 82].

We propose a learning framework where, given the Nash equilibrium action of the games, we jointly infer the graph structure that represents the interaction network as well as the individual marginal benefits. We further develop a second framework by considering the homophilous effect of individual marginal benefits in the interaction network. The first framework involves solving a convex optimization problem, while the second leads to a non-convex one for which we develop an algorithm based on alternating minimization. We test the performance of the proposed algorithms in inferring graph structures for network games and show that it is superior to the baseline approaches of sample correlation and regularized graphical Lasso [127], albeit developed for slightly different learning settings.

The main contributions of this section of my thesis are as follows. First, the proposed learning frameworks, to the best of our knowledge, are the first to address the problem of learning the graph structure of the broad class of network games with linear-quadratic payoffs and continuous actions. Second, our framework also allows for the inference of marginal benefits of the players which permits a range of applications such as target interventions. Third, we analyze several factors in the quadratic games that affect the learning performance, such as the strength of strategic complements or substitutes, the topological characteristics of the networks, and the homophilous effect of individual marginal benefits. Overall, this study constitutes a theoretical contribution to the studies of network games and may shed light on the understanding of strategic interactions in a wide range of practical scenarios, including business, education, governance, and technology adoption.

5.2 Network games of linear-quadratic payoffs

Consider a network of N individuals represented by a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote the node and edge sets, respectively. For any pair of individuals i and j, $G_{ij} = G_{ji} = 1$ if $(i, j) \in \mathcal{E}$ and $G_{ij} = G_{ji} = 0$ otherwise, where G_{ij} is the ij-th entry of the adjacency matrix $\mathbf{G}$. In a network game of linear-quadratic payoffs, an individual i chooses her action a_i to maximize her payoff, u_i, which has the following

form [21, 50, 82, 5]:

$$u_i = b_i a_i - \frac{1}{2} a_i^2 + \beta a_i \sum_{j \in \mathcal{V}} G_{ij} a_j. \tag{5.1}$$

In Eq. (5.1), the first term is contributed by i's own action where the parameter b_i is called the marginal benefit, and the third term comes from the peer effect weighted by the actions of her neighbors. The parameter β captures the nature and the strength of such peer effect: if $\beta > 0$, actions are called strategic complements; and if $\beta < 0$, actions are called strategic substitutes.

The quadratic game with payoff function in Eq. (5.1) represents a broad class of games that have been extensively studied in the literature, and has a number of desirable properties. First, it naturally allows for continuous actions (i.e., a_i is considered to be continuous); second, it can be used for modelling games of both strategic complements and substitutes, i.e., positive and negative spillover effect; third, it may also be used to approximate games with complex non-linear payoffs. For these reasons, games of linear-quadratic payoffs have been used to analyse crime activity, educational outcome, firm cooperation, and urban dynamics just to name a few [110].

One important advantage of the game in Eq. (5.1) is that it allows for an explicit solution for equilibrium behavior as a function of the network. To see this, let us define the vectorial forms $\mathbf{a} = [a_1, a_2, \cdots, a_N]^T$, $\mathbf{b} = [b_1, b_2, \cdots, b_N]^T$, and $\mathbf{u} = [u_1, u_2, \cdots, u_N]^T$, where we use the convention that the subscript i indicates the i-th entry of the vector. Taking the first-order derivative of the payoff u_i with respect to the action a_i in Eq. (5.1), we have:

$$\frac{\partial u_i}{\partial a_i} = b_i - a_i + \beta(\mathbf{G}\mathbf{a})_i. \tag{5.2}$$

Combining Eq. (5.2) for all i, it is clear that the following relationship holds, as pointed out in [21], for any (pure strategy) Nash equilibrium action $\mathbf{a}$:

$$(\mathbf{I} - \beta \mathbf{G})\, \mathbf{a} = \mathbf{b}, \tag{5.3}$$

hence

$$\mathbf{a} = (\mathbf{I} - \beta\mathbf{G})^{-1}\,\mathbf{b}, \tag{5.4}$$

where $\mathbf{I} \in \mathbb{R}^{N\times N}$ is the identity matrix. We adopt the critical assumption that the spectral radius of the matrix $\beta\mathbf{G}$, denoted by $\rho(\beta\mathbf{G})$, is less than 1, which guarantees the inversion of Eq. (5.4). Furthermore, as proved in [21], this assumption also ensures the uniqueness and stability of the Nash equilibrium action $\mathbf{a}$.

The equilibrium action $\mathbf{a}$ can be rewritten as $\mathbf{a} = \sum_{p=0}^{\infty} \beta^p \mathbf{G}^p \mathbf{b}$, and therefore has the following interpretations. If $\mathbf{b}$ is the all-one vector, then each entry of $\mathbf{a}$ is the Katz-Bonacich centrality [114, 43] of the corresponding node, i.e., the number of walks of any length p originated from that node discounted exponentially by β. As pointed out in [110], interestingly, this means despite the local neighborhood relationship in Eq. (5.1) the payoff interdependency actually spreads indirectly throughout the network. On the other hand, the formulation of Eq. (5.4) can also be interpreted as computing steady state opinions in studying opinion dynamics under a linear DeGroot model [71] and has been used in works on social network sensing [194].

From a different perspective, notice that $\mathbf{G}$ is a real and symmetric matrix hence has the following eigendecomposition: $\mathbf{G} = \boldsymbol{\chi}\boldsymbol{\Lambda}\boldsymbol{\chi}^T$. Plugging this into Eq. (5.4), the equilibrium action $\mathbf{a}$ can then be rewritten as $\mathbf{a} = \boldsymbol{\chi}(\mathbf{I} - \beta\boldsymbol{\Lambda})^{-1}\boldsymbol{\chi}^T\mathbf{b}$. Treating the marginal benefit $\mathbf{b}$ as a signal defined on the node set of the graph, the operation $\boldsymbol{\chi}^T\mathbf{b}$ can be interpreted as a Fourier-like transform for $\mathbf{b}$ according to the graph signal processing literature [158]. Given that the eigenvector associated with the largest/smallest eigenvalue of $\mathbf{G}$ is the most smooth/non-smooth hence corresponds to low-/high-frequency signal on the graph, the action $\mathbf{a}$ can thus be interpreted as a low-pass filtered version of $\mathbf{b}$ for $\beta > 0$, and a high-pass filtered version of it for $\beta < 0$. This matches our intuition that equilibrium action tends to be smooth on the interaction network for the case of strategic complements, and non-smooth for strategic substitutes.

5.3 Learning games with independent marginal benefits

Given the graph with an adjacency matrix $\mathbf{G}$, the marginal benefits $\mathbf{b}$, and the parameter β, Eq. (5.4) provides a way of computing $\mathbf{a}$, the Nash equilibrium action of the players. The graph structure, in many cases, can be naturally chosen from the application domain, such as a social or business network. However, these natural choices of graphs may not necessarily describe well the strategic interactions between the players, and a natural graph might not be easy to define at all in some applications. Compared to the underlying relationships captured by $\mathbf{G}$, it is often easier to observe the individual actions $\mathbf{a}$, such as the amount of effort committed by students in a joint course project, or the strategic moves made by firms in an industrial setting. In these cases, given the actions and the dependencies described in Eq. (5.1), it is therefore of considerable interest to infer the structure of the graph on which the game is played, hence revealing the hidden relationships between the players.

5.3.1 Learning framework

We consider N players, connected by a fixed interaction network $\mathbf{G}$, playing K different and independent games in each of which their payoffs depend not only on their own actions but also that of their neighbors. Let us define the marginal benefits for these K games as $\mathbf{B} = [\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \cdots, \mathbf{b}^{(k)}] \in \mathbb{R}^{N\times K}$, where each column of $\mathbf{B}$ is the marginal benefit vector for one game, and the corresponding actions of the players as $\mathbf{A} = [\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \cdots, \mathbf{a}^{(K)}] \in \mathbb{R}^{N\times K}$. We first consider in this section the case where, for each game, the marginal benefits of individual players follow independent and identical Gaussian distributions, and then addreee in Section 5.4 the dependent case. In our setting, the parameter that captures the strength of the network effect, β, can be either positive or negative, corresponding to strategic complements and strategic substitutes, respectively. Given the observed actions $\mathbf{A}$ and the parameter β, the goal is to infer a graph structure $\mathbf{G}$ as well as the marginal benefits $\mathbf{B}$, which best explain

$\mathbf{A}$ in terms of the relationship in Eq. (5.3).

To this end, we propose the following joint optimization problem of $\mathbf{G}$ and $\mathbf{B}$:

$$\begin{aligned} \min_{\mathbf{G},\mathbf{B}} \quad & f(\mathbf{G},\mathbf{B}) \\ & = ||(\mathbf{I}-\beta\mathbf{G})\mathbf{A}-\mathbf{B}||_F^2 + \theta_1||\mathbf{G}||_F^2 + \theta_2||\mathbf{B}||_F^2, \\ \text{s.t.} \quad & G_{ij} = G_{ji},\ G_{ij} \geq 0,\ G_{ii} = 0 \ \text{ for } \ \forall i,j \in \mathcal{V}, \\ & ||\mathbf{G}||_1 = N, \end{aligned} \tag{5.5}$$

where $\text{tr}(\cdot)$, $||\cdot||_F$, and $||\cdot||_1$ denote the trace operator, Frobenius norm, and element-wise L^1-norm of a matrix, respectively, and θ_1 and θ_2 are two non-negative regularization parameters. The first line of constraints ensures that $\mathbf{G}$ is a valid adjacency matrix, and the second constraint (the constraint on the L^1-norm) fixes the volume of the graph and permits to avoid trivial solutions. Without loss of generality the volume is chosen to be N. It is clear that, in the problem of Eq. (5.5), we aim at a joint inference of the graph structure $\mathbf{G}$ and the marginal benefits $\mathbf{B}$, such that the observed actions $\mathbf{A}$ are close to the Nash Equilibria of the K games played on the graph. The Frobenius norm on $\mathbf{G}$ is added as a penalty term to control the distribution of the edge weights of the learned graph (the off-diagonal entries of $\mathbf{G}$)[2], which, together with the L^1-norm constraint, bears similarity to the linear combination of L^1 and L^2 penalties in an elastic net regularization [210].

The effectiveness of the formulation in Eq. (5.5) depends on $\rho(\beta G)$. To see this, notice that under the assumption that $\mathbf{b}$ is IID Gaussian, the equilibrium action $\mathbf{a}$ follows a Gaussian distribution with covariance $(\mathbf{I}-\beta\mathbf{G})^{-2}$. If $\rho(\beta G)$ is close to zero, then $\mathbf{a}$ is almost independent from G, and it would be difficult to infer $\mathbf{G}$ from $\mathbf{a}$ in this scenario. On the other hand, if $\rho(\beta G)$ is close to one, the covariance is dominated by the eigenvector associated with the largest (when $\beta > 0$) or smallest (when $\beta < 0$) eigenvalue of $\mathbf{G}$. In this case, the action $\mathbf{a}$ clearly contains information about $\mathbf{G}$ which facilitates learning.

[2]Similar constraints have been adopted in [106, 72] for graph inference.

5.3.2 Learning algorithm

Given the non-negativity of G_{ij}, we can re-write the constraint: $||\mathbf{G}||_1 = \mathbf{1}^T\mathbf{G}\mathbf{1} = N$, where $\mathbf{1} \in \mathbb{R}^N$ is the all-one vector. The constraints in Eq. (5.5) therefore form a convex set. The problem of Eq. (5.5) is thus a quadratic program jointly convex in $\mathbf{B}$ and $\mathbf{G}$, and can be solved efficiently via the interior point methods [46]. In our experiments, we solve the problem of Eq. (5.5) using the Python software package CVXOPT [10]. In case of graphs of very large number of vertices, we can instead use operator splitting methods, e.g., alternating direction method of multipliers (ADMM) [45], to find a solution. The algorithm is summarized in Algorithm 2.

Algorithm 2: Learning games with independent marginal benefits

Input: Actions $\mathbf{A} \in \mathbb{R}^{N\times K}$ for K games, β, θ_1, θ_2
Output: Network $\mathbf{G} \in \mathbb{R}^{N\times N}$, marginal benefits $\mathbf{B} \in \mathbb{R}^{N\times K}$ for K games
Solve for $\mathbf{G}$ and $\mathbf{B}$ in Eq. (5.5)
return: $\mathbf{G}$, $\mathbf{B}$

5.4 Learning games with homophilous marginal benefits

A large number of studies in the literature of social sciences and economics have analyzed the phenomenon of homophily in social networks, which describes that individuals tend to associate and form ties with those that are similar to themselves [147, 109]. Since the marginal benefit vector $\mathbf{b}$ in each game can be thought of as the individual preferences toward a particular action, they may contribute, in the presence of the homophily effect, to the formation of the interaction network on which the game is played. The second formulation in my thesis is therefore to address the problem of learning games with such homophilous marginal benefits.

5.4.1 Learning framework

The homophily effect that is present in the marginal benefit vector $\mathbf{b}$ implies that $\mathbf{b}$ as a signal defined on the graph is relatively smooth, in the sense that nodes that are connected share similar marginal benefits. This may be quantified by the so-called Laplacian quadratic form on the graph:

$$\mathbf{b}^T\mathbf{L}\mathbf{b} = \frac{1}{2}\sum_{i,j\in\mathcal{V}} G_{ij}\left(b_i - b_j\right)^2, \tag{5.6}$$

where $\mathbf{L} = \text{diag}(\sum_{j\in\mathcal{V}} G_{ij}) - \mathbf{G}$ is the unnormalized (combinatorial) graph Laplacian matrix [67]. We therefore propose to replace the norm on $\mathbf{B}$ with this measure in the objective function of Eq. (5.5) to promote homophilous marginal benefits. This essentially assumes that the marginal benefits follow a multivariate Gaussian distribution with the precision matrix being the graph Laplacian. This leads to the following optimization problem:

$$\begin{aligned}
\min_{\mathbf{G},\mathbf{B}} \quad & h(\mathbf{G},\mathbf{B}) \\
=& ||(\mathbf{I}-\beta\mathbf{G})\mathbf{A}-\mathbf{B}||_F^2 + \theta_1||\mathbf{G}||_F^2 + \theta_2\ \text{tr}(\mathbf{B}^T\mathbf{L}\mathbf{B}), \\
\text{s.t.} \quad & G_{ij} = G_{ji},\ G_{ij} \geq 0,\ G_{ii} = 0 \quad \text{for} \quad \forall i,j \in \mathcal{V}, \\
& ||\mathbf{G}||_1 = N, \\
& \mathbf{L} = \text{diag}(\sum_{j\in\mathcal{V}} G_{ij}) - \mathbf{G},
\end{aligned} \tag{5.7}$$

where the third term in the objective is the sum of the Laplacian quadratic form for all the columns in $\mathbf{B}$, and the third constraint comes from the definition of the graph Laplacian $\mathbf{L}$. Like in Eq. (5.5), θ_1 and θ_2 are two non-negative regularization parameters. The problem of Eq. (5.7) is similar to that of Eq. (5.5), but with a different assumption that there exists the effect of homophily in the marginal benefits $\mathbf{b}$, whose strength is controlled by the regularization parameter θ_2, i.e., a larger θ_2 favors a stronger homophily effect, and vice versa.

5.4.2 Learning algorithm

Unlike the problem of Eq. (5.5), the problem of Eq. (5.7) is not jointly convex in $\mathbf{G}$ and $\mathbf{B}$ due to the third term in the objective function. We therefore adopt an alternating minimization scheme to optimize for the graph structure $\mathbf{G}$ and the marginal benefits $\mathbf{B}$ where, at each step, we solve for one variable while fixing the other.

Given $\mathbf{B}$, we first solve for $\mathbf{G}$ in Eq. (5.7). The constraints on $\mathbf{G}$ in Eq. (5.7) are the same as that in Eq. (5.5) and thus convex. Since $\theta_1 \geq 0$ and $\theta_2 \geq 0$, fixing $\mathbf{B}$ and solving for $\mathbf{G}$ results in a strongly convex objective, and consequently the problem admits a unique solution. We again solve this convex quadratic program using the package CVXOPT. Next, we fix $\mathbf{G}$ and solve for $\mathbf{B}$ in Eq. (5.7). By fixing $\mathbf{G}$, Eq. (5.7) becomes an unconstrained convex quadratic program, and thus admits a closed-form solution which can be obtained by setting the derivative to zero:

$$\frac{\partial h(\mathbf{G}, \mathbf{B})}{\partial \mathbf{B}} = -2\big((\mathbf{I} - \beta\mathbf{G})\mathbf{A} - \mathbf{B}\big) + 2\theta_2 \mathbf{L}\mathbf{B} = \mathbf{0}, \tag{5.8}$$

hence

$$\mathbf{B} = (\mathbf{I} + \theta_2 \mathbf{L})^{-1}(\mathbf{I} - \beta\mathbf{G})\mathbf{A}. \tag{5.9}$$

We iterative between the two steps until either the change in the objective $h(\mathbf{G}, \mathbf{B})$ is smaller than 10^{-4}, or a maximum number of iterations has been reached. This strategy is called block coordinate descent (BCD) and, since both subproblems are strongly convex, is guaranteed to converge to a local minimum (see Proposition 2.7.1 in [35]). The complete algorithm is summarized in Algorithm 3.

5.5 Experiments on synthetic data

In this section, we evaluate the performance of the proposed learning frameworks on synthetic networks that follow three types of random graph models, i.e., the Erdős–Rényi (ER), the Watts-Strogatz (WS), and the Barabási-Albert (BA) models. In the ER graph, an edge is created with a probability of $p = 0.2$ independently from all other possible edges. In the WS graph, we set the average degree of the vertices

Algorithm 3: Learning games with homophilous marginal benefits

Input: Actions $\mathbf{A} \in \mathbb{R}^{N\times K}$ for K games, β, θ_1, θ_2
Output: Network $\mathbf{G} \in \mathbb{R}^{N\times N}$, marginal benefits $\mathbf{B} \in \mathbb{R}^{N\times K}$ for K games
Initialize: $\mathbf{B}_0(:,k) \sim \mathcal{N}(\mathbf{0}, \mathbf{L}^\dagger)$
for $k = 1, \cdots, K$, $t = 1$, $\Delta = 1$ $\Delta \geq 10^{-4}$ and $t \leq$ # iterationsSolve for $\mathbf{G}_t$ in Eq. (5.7) given $\mathbf{B}_{t-1}$
Compute $\mathbf{L}_t$ using $\mathbf{G}_t$
$\mathbf{B}_t = (\mathbf{I} + \theta_2 \mathbf{L}_t)^{-1}(\mathbf{I} - \beta \mathbf{G}_t)\mathbf{A}$
$\Delta = |h(\mathbf{G}_t, \mathbf{B}_t) - h(\mathbf{G}_{t-1}, \mathbf{B}_{t-1})|$ (for $t > 1$)
$t = t + 1$
return: $\mathbf{G} = \mathbf{G}_t, \mathbf{B} = \mathbf{B}_t$.

to be $k = \lfloor \log_2(N) \rfloor$, with a probability of $p = 0.2$ for the random rewiring process. Finally, in the BA graph, we add $m = 1$ new node at each time by connecting it to an existing node in the graph via preferential attachment. All the graphs have $N = 20$ vertices in our experiments. Once the graphs are constructed, we compute $\beta > 0$ such that the spectral radius, $\rho(\beta \mathbf{G})$, varies between 0 and 1 hence satisfying the assumption in Section 5.2.

We adopt two different settings, one for generating the independent marginal benefits $\mathbf{b}$ and the other for the homophilous $\mathbf{b}$. In the independent case of Section 5.3, for each game, we generate realizations by considering $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In the homophilous setting of Section 5.4, we generate realizations by considering $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{L}^\dagger)$, where $\mathbf{L}^\dagger$ is the Moore-Penrose pseudoinverse of the groundtruth graph Laplacian $\mathbf{L}$. In both cases we further add Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \frac{1}{10}\mathbf{I})$ to the simulated marginal benefits. Now, given $\mathbf{b}$ and β, we compute the players' Nash equilibrium action $\mathbf{a}$ according to Eq. (5.4). We consider $K = 50$ games for each of which we generate the action $\mathbf{a}$.

We apply Algorithm 2 and Algorithm 3 to the respective settings to infer graph structures and compare against the groundtruth ones in a scenario of binary classification, i.e., either there exists an edge between i and j (positive case), or not (negative case). Since the ratio of positive cases is small for all the three types of graphs, we use the area under the curve (AUC) for the evaluation of the learning performance. We compare our algorithms with two baseline methods for inferring graph structures

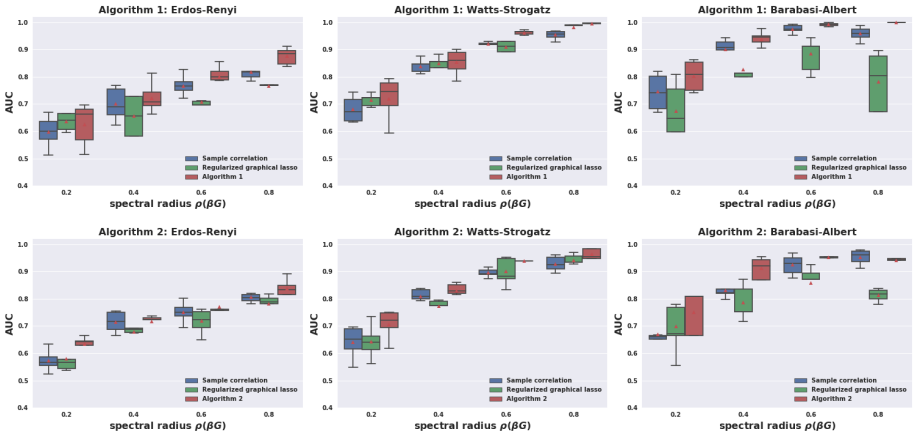


Figure 5-1: **Performance of the proposed algorithm and baselines in the setting of independent (top) and homophilous (bottom) marginal benefits.** The red triangle, the middle line, lower and upper boundaries of the box (interquartile range or IQR) correspond to mean, median, and 25/75 percentile of the data, respectively. The lower and upper whiskers extend maximally 1.5 times of IQR from 25 percentile downwards and 75 percentile upwards, respectively.

given data observations: the sample correlation and the regularized graphical Lasso in [127]. In the former case we consider the correlations between each pair of variables as "edge weights" in a learned graph, while in the latter case a graph adjacency matrix is computed as in our algorithms.

Notice that in the synthetic experiments we focus on the case of strategic complements, i.e., $\beta > 0$, to facilitate a fair comparison with the two baselines that only apply to this case. Our methods therefore also have the unique advantage of dealing with the case of strategic substitutes, i.e., $\beta < 0$.

5.5.1 Comparison of learning performance

The performance of the three methods in comparison is shown in Fig. 5-1 (top) for the case of independent marginal benefits. For Algorithm 2 and regularized graphical Lasso, we report the results using the parameter values that give the best average performance over 20 randomly generated graph instances[3]. First, we see that the performance of all the three methods increases with the spectral radius $\rho(\beta\mathbf{G})$ for

[3]Analysis of robustness of performance against regularization parameters is presented in Supplementary Material.

the majority of the cases. This pattern indicates that stronger strategic dependencies between actions of potential neighbors reveal more information about the existence of the corresponding links. Indeed, as $\rho(\beta\mathbf{G})$ increases, the action matrix $\mathbf{A}$ contains more information about the graph structure as explained in Section 5.3.1. Second, the performance of the proposed Algorithm 2 generally outperforms the two baselines in terms of recovering the locations of the edges of the groundtruth. Notice that for regularized graphical Lasso, the performance drops with larger value for $\rho(\beta\mathbf{G})$. One possible explanation is that, as $\rho(\beta\mathbf{G})$ becomes close to 1, the smallest eigenvalue of $\mathbf{I} - \beta\mathbf{G}$ approaches 0 resulting in a large ratio between the smallest and the largest eigenvalues of the empirical covariance of $\mathbf{a}$, which may lead to inaccurate estimation of the precision matrix in the graphical Lasso. In comparison, our method does not seem to be affected by such phenomenon. Finally, the performance of all the methods for the WS and BA graphs is generally better than that of the ER graphs, possibly because there exists more structural information in the former models than the latter.

The same results for the case of homophilous marginal benefits are shown in Fig. 5-1 (bottom). We observe the same increase in performance as $\rho(\beta\mathbf{G})$ increases for all the three methods, as well as the drop in performance towards large $\rho(\beta\mathbf{G})$ for regularized graphical Lasso. The proposed Algorithm 3 generally achieves superior performance in this scenario, which is expected due to the way the observations $\mathbf{A}$ are generated taking into account the regularization term in the objective in Eq. (5.7) that enforces homophily.

5.5.2 Learning performance with respect to different factors in network games

We now examine the performance of Algorithm 3 with respect to a number of factors, including the number of games, the noise intensity, the structure of the groundtruth network, and the strength of the homophily effect in marginal benefits (in Supplementary Material). The same results for Algorithm 2 are presented in Supplementary Material.

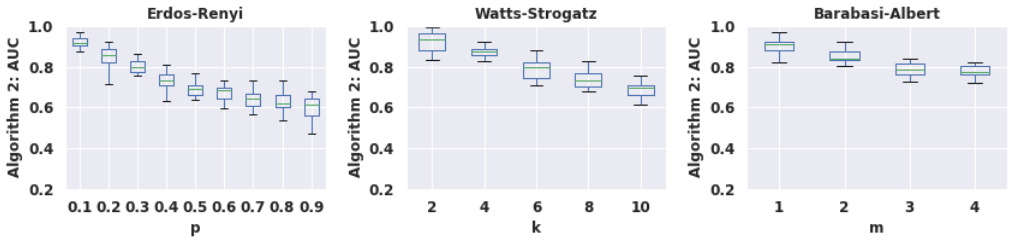


Figure 5-2: **Performance of Algorithm 3 versus structural properties of the network.**

Number of games. We are first interested in understanding the influence of the number of games K on the learning performance. In the following and all subsequent analyses, we choose $\rho(\beta\mathbf{G}) = 0.6$, and fix the parameters in Algorithm 3 to be the ones that lead to the best learning performance. In Fig. 5-5 (top), we vary the number of games and evaluate its effect on the performance. We see that in general, the performance of the algorithm increases, as more observations become available. The benefit is least obvious for the ER graph, suggesting that adding more observations does not help as much in improving the performance in this case when the edges in the graph appear more randomly.

Noise intensity in the marginal benefits. We now analyze the robustness of the result against noise intensity in the marginal benefits. With more noise in the marginal benefits, the observed actions $\mathbf{A}$ becomes noisier as well, hence possibly affecting the learning performance. As shown in Fig. 5-5 (bottom), the learning performance generally decays as the intensity of noise increases, which is expected. The performance of the model is relatively stable until the standard deviation of the noise becomes larger than 1.

Network structure. The random graphs used in our experiments have parameters that may affect the performance of the proposed algorithms. We therefore analyze the effect of p in the ER, k in the WS, and m in the BA graphs on the learning performance of the proposed algorithm. The larger these parameters, the higher the edge density in these random graph models. As shown in Fig. 5-2, the density of edges has a substantial effect on the learning performance for all the networks, i.e., the denser the edges, the worse the performance. One possible explanation is that,

in a sparse network the correlations between individuals' actions might contain more accurate information about the existence of dependencies hence edges between them, while in a dense network the influence from one neighbor is often mingled with that from another, which makes it more challenging to uncover pairwise dependencies.

5.5.3 Learning the marginal benefits

In our framework, we jointly infer the graph structure and the marginal benefits of the players. This is one of the main advantages of our algorithms, since the inference of marginal benefits can be critical for targeting strategies and interventions [82]. To test the performance of learning marginal benefits, for each random graph model, we generate a network with 20 nodes and simulate 50 games with $\rho(\beta\mathbf{G}) = 0.6$, for both independent and homophilous marginal benefits. We repeat this process for 30 times, and report the average performance of learning the marginal benefits in Table 5.1. The performance is measured in terms of the coefficients of determination (R^2), by treating the groundtruth and learned marginal benefits (both in vectorized form) as dependent and independent variables, respectively. As we can see, in both cases the R^2 values are above 0.9, which indicates that the learned marginal benefits are very similar to the groundtrith ones.

Table 5.1: **Performance (R^2) of learning marginal benefits.**

	Algorithm 2		Algorithm 3	
	mean	std	mean	std
ER graph	0.959	0.005	0.982	0.002
WS graph	0.955	0.007	0.921	0.010
BA graph	0.937	0.008	0.909	0.010

5.6 Experiments on real world data

The strategic interactions between players in real world situations may follow the formulation of the network games. Given this broad assumption, we present three examples of inferring the network structure in quadratic games in practical scenarios. The two examples in this section cover the inference of social network and trade network;

a third example on the inference of political network is presented in Supplementary Material.

5.6.1 Social network

We first consider inferring a social network between households in a village in rural India [23]. In particular, following the setting in [23], we consider the actions of each household as choosing the number of rooms, beds, and other facilities in their houses. The assumption is that there may exist strategic interactions between these households regarding constructing such facilities. In particular, when deciding to adopt new technologies or innovations, people have an incentive to conform to the social norms they perceive [204, 151], which are formed by the decisions made by their neighbors. For example, if neighbors adopt a specific facility, villagers tend to gain higher payoff after adopting the same facility by complying with social norms (i.e., strategic complements).

We consider each action as a strategy in a quadratic game, and we have 31 games with discrete actions made by 182 households. We then apply the proposed algorithms to infer the relationships between these households, and compare against a groundtruth network of self-reported friendship. Since we do not observe β, we treat it as a hyperparameter, and tune it within the range of $\beta \in [-3, 3]$. It can be seen from Table 5.2 that both of the proposed methods outperform regularized graphical Lasso by about 2.5% and sample correlation by about 10.7%[4], indicating that they can recover a social network structure closer to the groundtruth.

5.6.2 Trade network

We now consider inferring the global trade network. Specifically, we consider the overall trading activities of 235 countries on 96 export products and 96 import products

[4]The improvement is calculated by the absolute improvement in AUC normalized by the room for improvement. The best performance of Algorithm 2 is obtained with $\beta = 0.1$, $\theta_1 = 2^{-8.5}$, and $\theta_2 = 2^1$, while that of Algorithm 3 is obtained with $\beta = 2.6$, $\theta_1 = 2^7$, and $\theta_2 = 2^{-5.5}$. The positive sign of β in both cases indicates a strategic complement relationship between the households, which is consistent with our hypothesis.

in year 2008 as our observed actions[5]. This leads to 192 games (for both import and export actions) played by 235 agents (countries). By applying the proposed algorithms, we infer the relationships among nations regarding their strategic trading decisions and compare against a groundtruth which is the trading network in year 2002[6]. In constructing the groundtruth, we consider the edge weight between each pair of nations as the logarithmic of the total amount of trades (import plus export) between the two nations.

In the groundtruth trade network, each nation is connected with the ones with which it traded in 2002. This implies that the nation has different demand and supply compared to its neighbors, and their import and export actions tend to be different in the near future. Therefore, we expect a strategic substitute relationship between the nations when looking at their import and export activities in 2008.

We tune β within the range of $\beta \in [-1, 1]$. Table 5.2 shows that Algorithm 2 and Algorithms 2 outperform regularized graphical Lasso by 12.09% and 24.85%, respectively[7]. The larger performance gain in this case is due to the fact that both sample correlation and regularized graphical Lasso are suitable only for strategic complement and not strategic substitute relationships. Furthermore, Algorithm 3 performs better than Algorithm 2 in this example, which implies a homophilous distribution of marginal benefits across neighboring nations.

Table 5.2: **Performance (AUC) of learning the structure of the social network and the trade network.**

	Social network	Trade network
Sample correlation	0.525	0.523
Regularized graphical Lasso	0.564	0.570
Algorithm 2	0.575	0.622
Algorithm 3	0.576	0.677

[5]Data can be accessed via `https://atlas.media.mit.edu/en/resources/data/`. The trading activities are classified by the 2002 edition of the HS (Harmonized System).

[6]The trading network from previous years provides a foundation for nations to make decisions and thus can be thought of as a groundtruth. The year 2002 is the latest year before 2008 for which trading data are available.

[7]The best performance of Algorithm 2 is obtained with $\beta = -0.6$, $\theta_1 = 2^1$, and $\theta_2 = 2^{-10}$, and that of Algorithm 3 is obtained with $\beta = -0.7$, $\theta_1 = 2^{11.5}$, and $\theta_2 = 2^{-15.5}$. The negative sign of β in both cases indicates a strategic substitute relationship between the nations, which is consistent with our hypothesis.

5.7 Discussion

In this section of my thesis, we have proposed two novel learning frameworks for a joint inference of graph structure and individual marginal benefits for a broad class of network games, i.e., games with linear-quadratic payoffs. We believe that the present project may shed light on the understanding of network games (in particular those with linear-quadratic payoffs), and contribute to the vibrant literature of learning hidden relationships from data observations.

The proposed approaches can benefit a wide range of practical scenarios. For instance, the learned graph, which captures the strategic interactions between the players, may be used for detecting communities formed by the players [78]. This can, in turn, be used for purposes such as stratification. Another use case is to compute centrality measures of the nodes in the network, which may help in designing efficient targeting strategies in marketing scenarios [135]. Finally, the joint inference of the graph and the marginal benefits can help a central planner who wishes to design intervention mechanisms achieve specific planning objectives. One such objective could be the maximization of the total payoffs of all players, which can be done by adjusting, according to the network topology, the marginal benefits via incentivization [82]. Another objective could be the reduction of inequality between the players in terms of their payoffs, which can be done by adjusting network topology via encouraging the formation of certain new relationships.

There remain many interesting directions to explore. For example, building upon the promising empirical results presented in this section of my thesis, it would be important to study theoretical guarantees of the proposed algorithms in recovering the graph structure. It would also be interesting to consider graph inference given partial or incomplete observations of the actions, especially in the case where it is costly to observe the actions of all the network players, or consider a setting where the underlying relationships between the players may evolve over time, which can be modeled by dynamic graph topologies. Finally, the inference framework may need to be adapted accordingly for network games of different payoff functions. We leave

these studies as future work.

Supplementary Material

Robustness against regularization parameters

We analyze the robustness of the performance of Algorithm 2 against the regularization parameter θ_1 in Eq. (5.5), and the results averaged over 20 random graph instances are presented in Fig. 5-3. In general, in addition to the effect of $\rho(\beta\mathbf{G})$ discussed in the main text, we see a consistent pattern across the three graph models that link the values of θ_1 and θ_2 to the learning performance. Specifically, when θ_1 is smaller than around 10^2, there is a region where a certain ratio of θ_1 to θ_2 leads to optimal performance, suggesting that in this case, the second and third terms are the dominating factors in the optimization of Eq. (5.5). A phase transition takes place when θ_1 is larger than 10^2, where the performance becomes largely constant. The reason behind this behavior is as follows. When θ_1 increases, the Frobenius norm of $\mathbf{G}$ in the objective function of Eq. (5.5) tends to be small. Given a fixed element-wise L^1-norm of $\mathbf{G}$, this leads to a more uniform distribution of the off-diagonal entries. When θ_1 is large enough, the edge weights become almost the same, leading to a constant AUC measure.

Similarly, we present in Fig. 5-4 the performance of Algorithm 3 with respect to different values of θ_1 and θ_2 in Eq. (5.7). We see that the patterns are generally consistent with that in Fig. 5-3, with one noticeable difference being that there also seems to be a phase transition taking place around the value of 10^{-1} for θ_2. One possible explanation for this behavior is that, when θ_2 is large enough, the trace term in the objective function of Eq. (5.7) tends to be small, making the resulting graph with fewer edges but with larger weights. This contributes to an AUC score that is mostly constant.

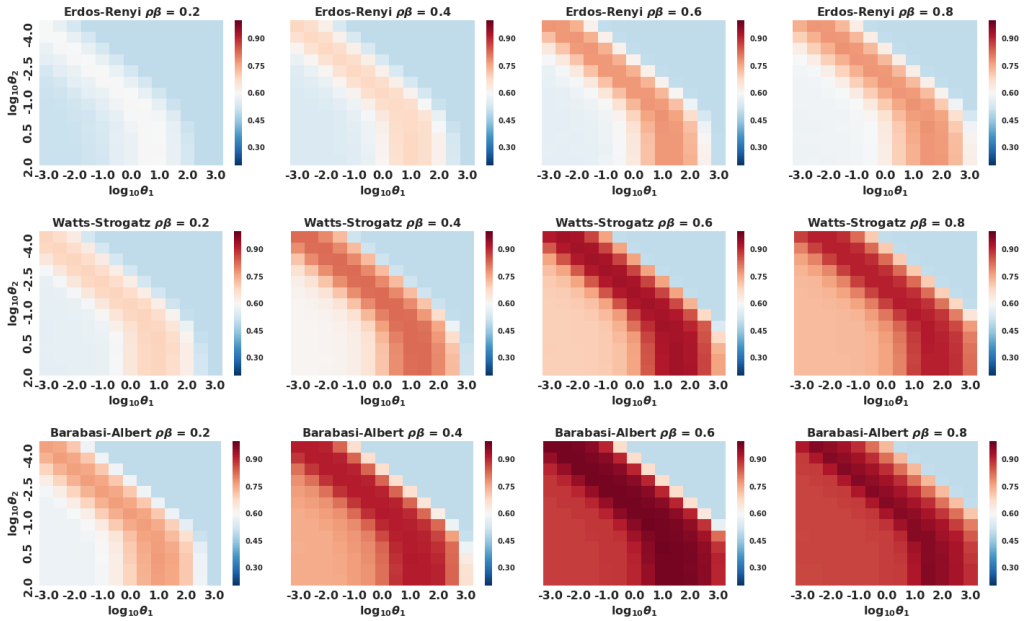


Figure 5-3: **Performance (AUC) of Algorithm 2 with respect to $\rho(\beta\mathbf{G})$, θ_1, and θ_2.**

Performance of Algorithm 3 with respect to number of games and noise intensity in marginal benefits

The performance of Algorithm 3 with respect to the number of games and noise intensity in marginal benefits analysed in Section 5.5.2 is presented in Fig. 5-5.

Performance of Algorithm 3 with respect to strength of homophily effect

We analyze the influence of the strength of homophily on the learning performance of Algorithm 3. We consider three scenarios, i.e., weak, medium and strong homophily effect. To this end, we generate the marginal benefits $\mathbf{b}$ as linear combinations of the eigenvectors corresponding to the 1^{st}-5^{th}, 6^{th}-10^{th}, and 11^{th}-15^{th} smallest eigenvalues of the graph Laplacian. Due to the properties of the eigenvectors, these three sets lead to different quantities for the Laplacian quadratic form, hence corresponding to weak, medium and strong homophily effect, respectively. Notice that the presence of the homophily effect in $\mathbf{B}$ tends to imply homophily in $\mathbf{A}$ for the following reason.

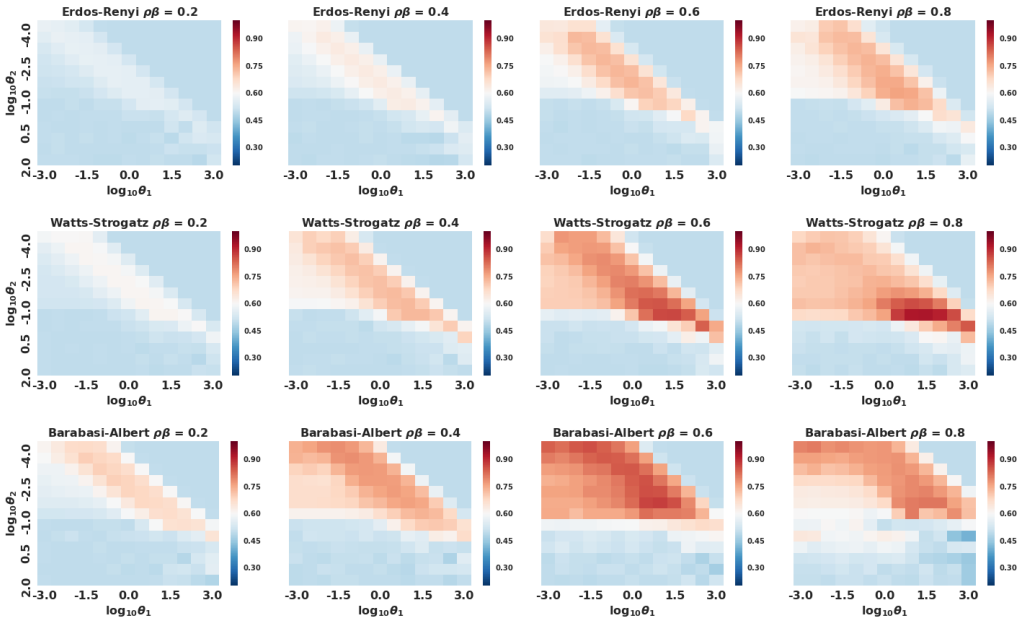


Figure 5-4: **Performance (AUC) of Algorithm 3 with respect to $\rho(\beta\mathbf{G})$, θ_1, and θ_2.**

Regardless of the characteristics of the game, a higher marginal benefit **b** is more likely to incentivize higher activity level **a** due to the first term of the payoff function in Eq. (5.1). Therefore, homophily in **B** tends to lead to homophily in **A**, hence revealing more information about the graph structure. As shown in Fig. 5-6, for all the three types of networks, the stronger the homophily in the marginal benefits, the better the learning performance.

Results in Section 5.5.2 for Algorithm 2

The performance of Algorithm 2 with respect to the factors analysed in Section 5.5.2 is presented in Fig. 5-7.

Inference of political network

The third real world example we considered is the inference of the relationship between the cantons in Switzerland in terms of their political preference. To this end, we consider voting statistics from the national referendums for 37 federal initiatives in

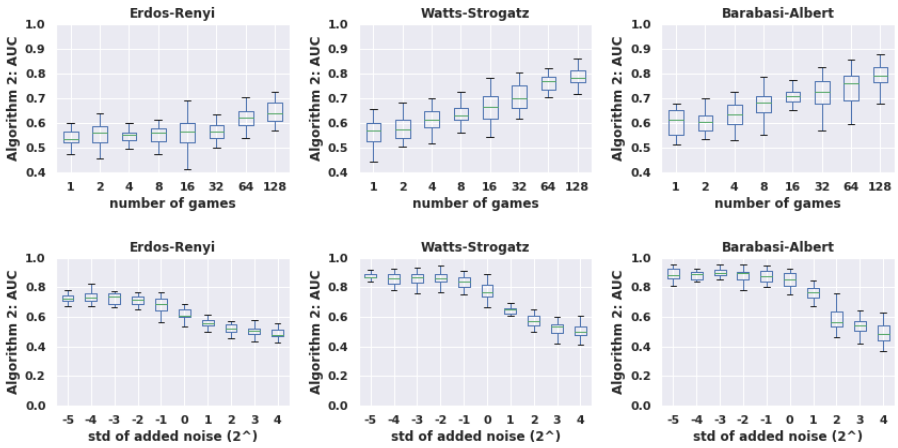


Figure 5-5: **Performance of Algorithm 3 versus number of games (top) and noise intensity in marginal benefits (bottom).**

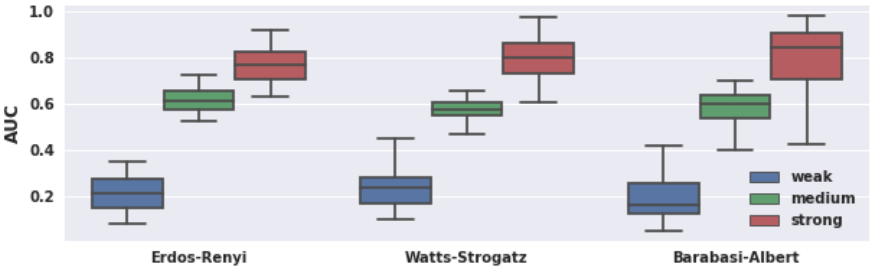


Figure 5-6: **Performance of Algorithm 3 versus strength of homophily in the marginal benefits.**

Switzerland between 2008 and 2012[8]. Specifically, we consider the percentage of voters supporting each initiative in the 26 Swiss cantons as the observed actions. This leads to 37 games (initiatives) played by 26 agents (cantons). By applying the proposed algorithms, we infer a network that captures the strategic political relationship between these cantons reflected by their votes in the national referendums[9].

Unlike the previous examples, it is more difficult to define a groundtruth network in this case. Instead, we apply spectral clustering [193] to the learned network and interpret the obtained clusters of cantons. The three-cluster partition of the networks learned by Algorithm 2 and Algorithm 3 are presented in Fig. 5-8(a) and Fig. 5-8(b),

[8]The voting statistics were obtained via `http://www.swissvotes.ch`.

[9]We tune β within the range of $[-1, 1]$. For Algorithm 2 we report results with $\beta = 0.6$, $\theta_1 = 2^{-6.2}$, and $\theta_3 = 2^{-1.65}$. For Algorithm 3 we report results with $\beta = 0.67$, $\theta_1 = 2^2$, and $\theta_2 = 2^3$. The positive sign of β in both cases indicates a strategic complement relationship between the cantons.

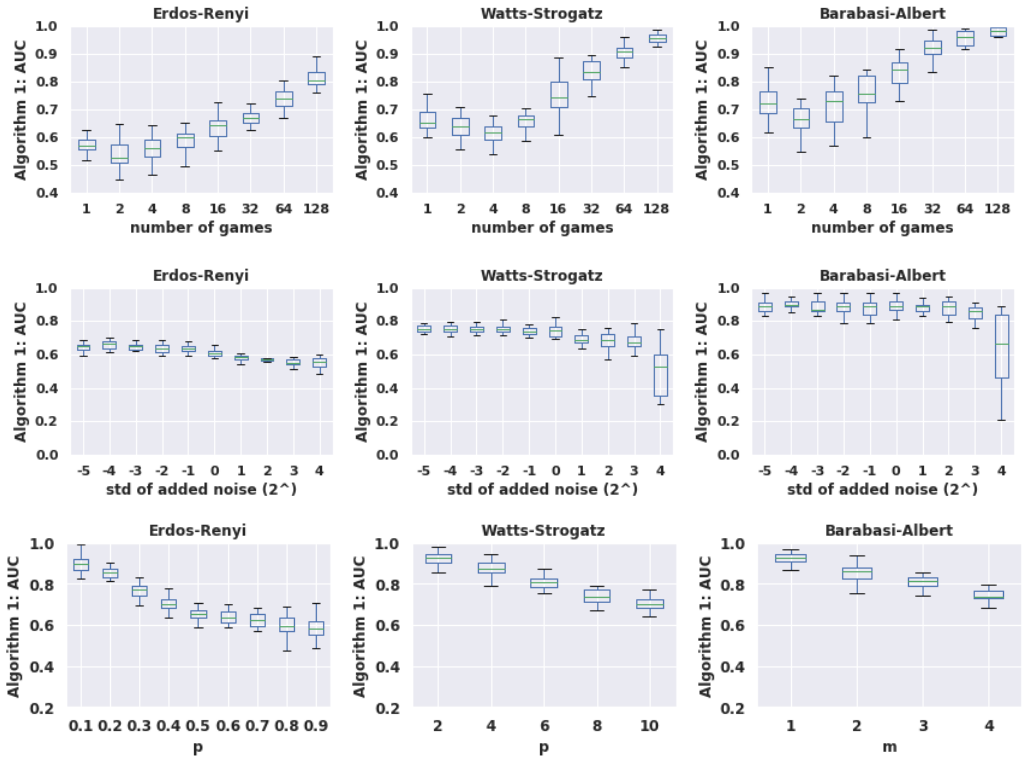


Figure 5-7: **Performance of Algorithm 2 versus number of games (top), noise intensity in marginal benefits (middle), and structural properties of the network (bottom).**

respectively. As we can see, the clusters obtained in the two cases are largely consistent, with the blue and yellow clusters generally corresponding to the French-speaking and German-speaking cantons, respectively. The red cluster, in both cases, contains the five cantons of Uri, Schwyz, Nidwalden, Obwalden and Appenzell Innerrhoden, which are all considered among the most conservative ones in Switzerland. This demonstrates that the learned networks are able to capture the strategic dependence between cantons within the same cluster, which tend to vote similarly in national referendums.

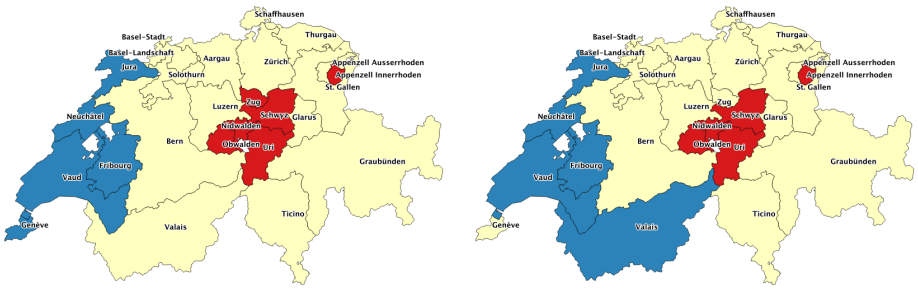


Figure 5-8: **Clustering of Swiss cantons based on the political network learned by Algorithm 2 (left) and Algorithm 3 (right).**

Chapter 6

Conclusion

My thesis is motivated by the intricate linkage between individuals' characteristics, actions, and their networks. People form network connections based on homophily; individuals' networks also shape their actions. Pervasive behavioral data provides opportunities for a richer view of the decisions on networks. Yet, the increasing volume, complex structures, and dynamics of behavioral data stretch the limit of conventional methods. I aspire to bridge mathematical modeling (i.e., machine learning, game theory, and network science) and computational social science to understand human behaviors on networks.

6.1 Summary

My thesis thus far have tackled this problem from four directions.

1. I developed frameworks to **learn the hidden network connections** based on individuals' decisions.

2. I empirically investigated and modeled **how social influence spread over networks**.

3. I studied how to **leverage influential nodes for selective network interventions**.

4. I developed methods to **incorporate network and other complex data structures** for inference problems on networks, i.e., recommendations and counterfactual predictions.

These four areas jointly support my research agenda to (1) leverage social interactions to understand human behavior; and (2) develop computational tools for behavioral predictions, causal inference, and network inference. In what follows, I summarize the thesis.

1. Learning the network structure from decisions In many social settings, social connections are either unobserved or noisily measured. Individual actions provide information about the underlying interaction structures due to the dependencies of neighbors' actions. Jointly with Xiaowen Dong, Junfeng Wu, and Alex Pentland [182], we formalized this idea with a linear-quadratic network game. This game is an approximation of all static games with continuous utility functions. We used Nash Equilibrium to approximate users' actions by assuming that rational agents maximize their utilities. We provided conditions under which network structure can be inverted from observed actions, and we performed several empirical applications of the framework in the paper, including 1) inferring political alliances from voting outcomes in national referendums for 37 federal initiatives in Switzerland; 2) inferring a global trade network of 235 countries based on import and export activities; 3) inferring the social network based on a range of household decisions in rural India. We are currently working on the generalization of this framework to tackle a wide range of games. In particular, we built a deep auto-encoder framework, in which we used the encoder to infer the underlying connections and the decoder to proxy the decision-making.

2. How social influence spread over networks Several empirical studies have shown that social influence propagates beyond direct neighbors in relatively costless online decision-making settings. Yet, precisely how influence plays a role in costly offline behaviors and spreads through a social network remains unclear. Jointly with

Xiaowen Dong, Matias Travizano, Esteban Moro, and Alex Pentland, we leveraged the high-resolution mobile phone data and a new behavioral matching framework to study how social influence propagates and affects individual offline behavior [**?**]. Our results showed that propagation within the network persists in shaping individual decisions through up to three degrees of separation in two non-routine offline environments. We also found that exposure to adoption behavior does not sufficiently explain this social influence's ripple effect. Therefore, we proposed a Bayesian learning model based on local information aggregation, which better predicts individual adoption behavior than exposure-based contagion models. This means that the local information aggregation is a paramount ingredient for understanding the diffusion of influence, and it could have implications in marketing and political campaigns, such as developing new centrality measures. Part of this work in included in the book "Spreading Dynamics in Social Systems."

In a project with Tara Sowrirajan and Alex Pentland, we further enriched the diffusion model with user characteristics and the latent communities. For example, users may be inclined to copy the decisions of similar others, while making different decisions to those in different social groups. Motivated by this, we proposed a model to jointly infer the endogenous network formation and the adoption decisions affected by influence varying across different empirically-identified communities [187]. The results from two empirical studies revealed the social dynamics among hidden communities and enabled us to infer influential social groups by combining with the socio-demographic data.

3. Leveraging influential nodes for selective network interventions Existing centrality measures study the connectedness of individuals. However, these measures are less helpful in some applications where the objective is to target users who spread positive influence, such as viral marketing or political campaigns. In a joint study with Yehonatan Yella, Rodrigo Ruiz, and Alex Pentland in Chapter 4, we developed the "contextual centrality" to guide such applications. In particular, contextual centrality evaluates individuals' importance based on network positions and nodal characteristics.

It generalizes over existing centrality measures and provides insights on both local and global diffusion. Contextual centrality is shown to perform better in the empirical analysis and simulations on the marketing campaigns for microfinance and weather insurance in rural villages in India and China. This work provides building blocks for integrating network structures and note features in future network studies.

4. Leveraging social connections for the inference problems Social networks contain hidden information about users' preferences and characteristics. "Network embedding" is the technique to infer hidden node features from observed network structure. I built upon this technique to develop tools for the inference problems of human behaviors on the network.

Joint with Rodrigo Ruiz, Xiaowen Dong, and Alex Pentland, we applied network embedding to recommendation systems by developing a novel geometric deep learning approach. Relative to the prior state-of-the-art recommender systems, which employed either nodal characteristics or network structure—but not both—for the recommendation, our approach enables recommendation systems to combine both sources of information and predict individual preferences using data with a complex structure. We applied the methodology to Yelp review data and predicted customer preferences for restaurants they have not rated, utilizing information on historical ratings, socio-demographics, business characteristics, check-in information, geographical information, and social networks. The methodology has a wide range of other potential applications; besides Yelp, the paper also demonstrates how to apply the framework to recommendations over Douban.com and Netflix.

Moreover, the network information can be used to improve causal inference. The bias amplification literature shows that controlling for the instrument variables increases the estimation bias. However, recent studies on applying representation learning in counterfactual predictions do not distinguish bias amplifiers from other confounders. In a joint study with Martin Saveski, Dean Eckles, and Alex Pentland, we proposed a novel deep learning framework to deal with this issue. In particular, we used group lasso regularization to enforce that the learned representations are

highly associated with both the treatment and the outcome[186]. To demonstrate the effectiveness of this approach, we ran naturalistic simulations using the Facebook 100 data set and illustrated the potential of using network information for observational inference in general.

6.2 Future work

To conclude, my research agenda is to use large-scale data sets, network theory, and machine learning to understand human behavior over social networks. To continue pursuing this agenda, I am excited to investigate three directions in the future.

Machine learning and data mining over social networks The increasing volume and complex structures of behavioral datasets extend beyond the scope of existing methods. Hence, there is a need for methods integrating complicated data structures, including dynamics, spatial-temporal correlations, multi-modal structures, and multi-layer networks. Take mobility data as an example; how can we extract patterns when there exist intrinsically spatial-temporal correlations among observations of the traffic flow? In social settings, how can we effectively use the dynamic interactions and the evolving relationships to extract intrinsic characteristics of individuals? It is essential to incorporate psychological and sociological theories into modeling and machine learning to answer these challenging questions.

Machine learning and big data for social good Data science roots in practical applications. Existing abundant data provides an excellent opportunity to solve real-world problems. In the past, I have worked with (1) Andorra government in tourism analysis and transportation congestion with CDR and twitter data [184, 185]; (2) Chinese research institutions to monitor commuting patterns with CDR and detect congestion spots with bus GPS data [183]; and (3) to understand issues faced by the refugees. Going forward, I am strongly motivated to continue solving pressing societal problems with user behavioral modeling. These collaborations may yield future funding opportunities. Questions of interest include but not limited to: (1)

how does integrate with residents affect the well-being of refugees? (2) How to control social networks and change individuals' incentives to increase information accessibility?

Adaptive interventions on dynamic networks I also intend to utilize the analysis and modeling of human behaviors to guide adaptive interventions. There are rich topics to explore, especially accounting for the evolving and the connectedness of human behaviors. A broader topic is to design interventions that incorporate the following elements: (1) evolving of social networks (formation and breakup of social links); (2) social learning process that changes individuals' preferences and hence decisions. This is potentially a large-scale project that would require collaborations with different researchers and might require a diverse skill set, including theoretical modeling, empirical methods, and the capacity to run field experiments.

In the longer-term, I intend to continue focusing on behaviors on networks and solve socially-relevant problems. I am fortunate to have collaborated with researchers from diverse fields and am convinced that by pursuing these types of collaborations in the future, I will be able to address socially-impactful questions that stand at the intersection of social science, network science, and machine learning.

Bibliography

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*, 2017.

[2] Daron Acemoglu, Kostas Bimpikis, and Asuman Ozdaglar. Dynamics of information exchange in endogenous social networks. *Theoretical Economics*, 9(1):41–97, 2014.

[3] Daron Acemoglu, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

[4] Daron Acemoglu and Asuman Ozdaglar. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1(1):3–49, 2011.

[5] Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. Networks, shocks, and systemic risk. Working Paper 20931, National Bureau of Economic Research, February 2015.

[6] Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608, 2015.

[7] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[8] Mina Ameri, Elisabeth Honka, and Ying Xie. Word-of-mouth, observational learning, and product adoption: Evidence from an anime platform. 2016.

[9] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15. ACM, 2008.

[10] M. S. Andersen, J. Dahl, and L. Vandenberghe. CVXOPT: A Python package for convex optimization. pages 301–320, version 1.2.0. Available at cvxopt.org, 2018.

[11] Chris Anderson. *The long tail: Why the future of business is selling less of more.* Hachette Books, 2006.

[12] Joshua D Angrist. The perils of peer effects. *Labour Economics*, 30:98–108, 2014.

[13] Sinan Aral. Social science: Poked to vote. *Nature*, 489(7415):212–214, 2012.

[14] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[15] Sinan Aral and Christos Nicolaides. Exercise contagion in a global social network. *Nature communications*, 8:14753, 2017.

[16] Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science*, 57(9):1623–1639, 2011.

[17] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[18] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 33–42, 2012.

[19] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

[20] Coralio Ballester, Antoni Calvó-Armengol, and Yves Zenou. Who's who in networks. wanted: The key player. *Econometrica*, 74(5):1403–1417, 2006.

[21] Coralio Ballester, Antoni Calvó-Armengol, and Yves Zenou. Who's who in networks. Wanted: The key player. *Econometrica*, 74(5):1403–1417, 2006.

[22] A. Bandura and D. C. McClelland. Social learning theory. 1977.

[23] Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. The diffusion of microfinance. *Science*, 341(6144):1236498, 2013.

[24] Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. Gossip: Identifying central individuals in a social network. Technical report, National Bureau of Economic Research, 2014.

[25] Abhijit Banerjee, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. The miracle of microfinance? evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7(1):22–53, 2015.

[26] Abhijit V Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817, 1992.

[27] Abhijit V Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. Using gossips to spread information: Theory and evidence from two randomized controlled trials. 2017.

[28] Abhijit V Banerjee and Esther Duflo. The experimental approach to development economics. *Annu. Rev. Econ.*, 1(1):151–178, 2009.

[29] Jie Bao, Yu Zheng, David Wilkie, and Mohamed Mokbel. Recommendations in location-based social networks: A survey. *GeoInformatica*, 19(3):525–565, 2015.

[30] Adarsh Barik and Jean Honorio. Provable computational and statistical guarantees for efficient learning of continuous-action graphical games. *arXiv:1911.04225*, 2019.

[31] JM Bernardo MJ Bayarri, JO Bergen AP Dawid, D Heckerman AE M Smith, and M West. Hierarchical bayesian models for applications in information retrieval. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, page 25. Oxford University Press, 2003.

[32] Alexandre Belloni, Victor Chernozhukov, et al. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[33] Alexandre Belloni, Victor Chernozhukov, and Ying Wei. Honest confidence regions for a regression parameter in logistic regression with a large number of controls. Technical report, cemmap working paper, Centre for Microdata Methods and Practice, 2013.

[34] Dirk Bergemann and Deran Ozmen. Optimal pricing with recommender systems. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 43–51. ACM, 2006.

[35] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1995.

[36] Bruce J Biddle. Recent developments in role theory. *Annual review of sociology*, 12(1):67–92, 1986.

[37] Norbert Binkiewicz, Joshua T Vogelstein, and Karl Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 2017.

[38] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[39] Christopher M Bishop and Michael E Tipping. Bayesian regression and classification. *Nato Science Series sub Series III Computer And Systems Sciences*, 190:613–632, 2003.

[40] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[41] Per Block and Thomas Grund. Multidimensional homophily in friendship networks. *Network Science*, 2(2):189–212, 2014.

[42] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-based systems*, 46:109–132, 2013.

[43] P. Bonacich. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182, 1987.

[44] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295, 2012.

[45] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[46] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[47] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017.

[48] Yann Bramoullé and Rachel Kranton. Public goods in networks. *Journal of Economic Theory*, 135(1):478–494, 2007.

[49] Yann Bramoullé and Rachel Kranton. Games played on networks. *The Oxford Handbook of the Economics of Networks, Yann Bramoullé, Andrea Galeotti, and Brian Rogers, eds.*, 2016.

[50] Yann Bramoullé, Rachel Kranton, and Martin D'amours. Strategic interaction and networks. *American Economic Review*, 104(3):898–930, 2014.

[51] Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008.

[52] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[53] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and deep locally connected networks on graphs. In *International Conference on Learning Representations (ICLR)*, 2014.

[54] Robin Burke. Hybrid recommender systems: survey and experiments. *User Modeling and User-Adapted Interaction*, 2002.

[55] John W Byers, Michael Mitzenmacher, and Georgios Zervas. The groupon effect on yelp ratings: a root cause analysis. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 248–265. ACM, 2012.

[56] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637, 2018.

[57] Jing Cai, Alain De Janvry, and Elisabeth Sadoulet. Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108, 2015.

[58] Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.

[59] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

[60] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)*, 10(4):1, 2008.

[61] Yubo Chen, Qi Wang, and Jinhong Xie. Online social interactions: A natural experiment on word of mouth versus observational learning. *Journal of marketing research*, 48(2):238–254, 2011.

[62] Chen Cheng, Haiqin Yang, Irwin King, and Michael R Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. 2012.

[63] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

[64] Nicholas A. Christakis and James H. Fowler. *Connected: The surprising power of our social networks and how they shape our lives*. Little, Brown, 2009.

[65] Nicholas A Christakis and James H Fowler. Social network sensors for early detection of contagious outbreaks. *PloS one*, 5(9):e12948, 2010.

[66] Nicholas A Christakis and James H Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32(4):556–577, 2013.

[67] F. R. K. Chung. *Spectral graph theory*. American Mathematical Society, 1997.

[68] Jacob Cohen. Statistical power analysis for the behavioral sciences . hilsdale. *NJ: Lawrence Earlbaum Associates*, 2, 1988.

[69] Sergio Currarini, Matthew O Jackson, and Paolo Pin. Identifying the roles of race-based choice and chance in high school friendship network formation. *Proceedings of the National Academy of Sciences*, 107(11):4857–4861, 2010.

[70] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pages 3844–3852, 2016.

[71] M. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69:118–121, 1974.

[72] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23):6160–6173, 2016.

[73] X. Dong, D. Thanou, M. Rabbat, and P. Frossard. Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*, 36(3):44–63, 2019.

[74] Xiaowen Dong, Yoshihiko Suhara, Burcin Bozkaya, Vivek K. Singh, Bruno Lepri, and Alex Pentland. Social bridges in urban purchase behavior. *ACM Transactions on Intelligent Systems and Technology, Special Issue on Urban Intelligence*, 9(3):33:1–33:29, 2018.

[75] Wenjing Duan, Bin Gu, and Andrew B Whinston. Informational cascades and software adoption on the internet: an empirical investigation. *MIS quarterly*, pages 23–48, 2009.

[76] Dean Eckles and Eytan Bakshy. Bias and high-dimensional adjustment in observational studies of peer effects. *arXiv preprint arXiv:1706.04692*, 2017.

[77] Econsultancy and Monetate. The realities of online personalization, 2013.

[78] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[79] James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *Bmj*, 337:a2338, 2008.

[80] James H Fowler and Nicholas A Christakis. Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, page 200913149, 2010.

[81] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[82] Andrea Galeotti, Benjamin Golub, and Sanjeev Goyal. Targeting interventions in networks. *arXiv:1710.06026*, 2017.

[83] Chao Gao, Zongming Ma, Anderson Y Zhang, Harrison H Zhou, et al. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185, 2018.

[84] Vikas Garg and Tommi Jaakkola. Learning tree structured potential games. In *Advances in Neural Information Processing Systems*, pages 1552–1560, 2016.

[85] Vikas Garg and Tommi Jaakkola. Local aggregative games. In *Advances in Neural Information Processing Systems*, pages 5341–5351, 2017.

[86] Asish Ghoshal and Jean Honorio. From behavior to sparse graphical games: Efficient recovery of equilibria. In *Proceedings of the IEEE Allerton Conference on Communication, Control, and Computing*, 2016.

[87] Asish Ghoshal and Jean Honorio. Learning graphical games from behavioral data: Sufficient and necessary conditions. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1532–1540, 2017.

[88] Asish Ghoshal and Jean Honorio. Learning sparse polymatrix games in polynomial time and sample complexity. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 1486–1494, 2018.

[89] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *Proc. of the 28th Inter. Conf. on Machine Learning*, pages 561–568, Bellevue, Washington, USA, 2011.

[90] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proc. of the 16th ACM SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining*, pages 1019–1028, Washington, DC, USA, 2010.

[91] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):21, 2012.

[92] Manuel Gomez Rodriguez, Jure Leskovec, and Bernhard Schölkopf. Structure and dynamics of information pathways in online media. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 23–32. ACM, 2013.

[93] Cyril Goutte, Peter Toft, Egill Rostrup, Finn Å Nielsen, and Lars Kai Hansen. On clustering fmri time series. *NeuroImage*, 9(3):298–310, 1999.

[94] Sanjeev Goyal and José Luis Moraga-Gonzaléz. R&D networks. *Rand Journal of Economics*, 32(4):686–707, 2001.

[95] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[96] Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[97] Donald P Green and Alan S Gerber. *Get out the vote: How to increase voter turnout*. Brookings Institution Press, 2015.

[98] Yue Guan, Qiang Wei, and Guoqing Chen. Deep learning based personalized recommendation with multi-view information integration. *Decision Support Systems*, 2019.

[99] Roger Guimera, Stefano Mossa, Adrian Turtschi, and LA Nunes Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799, 2005.

[100] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[101] Michal Herzenstein, Utpal M Dholakia, and Rick L Andrews. Strategic herding behavior in peer-to-peer loan auctions. *Journal of Interactive Marketing*, 25(1):27–36, 2011.

[102] M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *ArXiv e-prints*, November 2011.

[103] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[104] Jean Honorio and Luis E Ortiz. Learning the structure and parameters of large-population graphical games from behavioral data. *Journal of Machine Learning Research*, 16:1157–1210, 2015.

[105] Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation. *Management Science*, 60(4):805–823, 2013.

[106] C. Hu, L. Cheng, J. Sepulcre, K. A. Johnson, G. E. Fakhri, Y. M. Lu, and Q. Li. A spectral graph regression model for learning brain connectivity of alzheimer's disease. *PLoS ONE*, 10(5):e0128136, May 2015.

[107] Mohammad Irfan and Luis Ortiz. On influence, stable behavior, and the most influential individuals in networks: A game-theoretic approach. *Artificial Intelligence*, 215:79–119, 2014.

[108] Raghuram Iyengar, Christophe Van den Bulte, and Jeonghye Choi. Social contagion in new product adoption: Networks versus co-location. 2010.

[109] Matthew O. Jackson. *Social and economic networks*. Princeton University Press, 2010.

[110] Matthew O. Jackson and Yves Zenou. Games on networks. *Handbook of Game Theory, Vol. 4, Peyton Young and Shmuel Zamir, eds.*, 2014.

[111] Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. Recommender systems—beyond matrix completion. *Communications of the ACM*, 59(11):94–102, 2016.

[112] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[113] Zsolt Katona, Peter Pal Zubcsek, and Miklos Sarvary. Network effects and personal influences: The diffusion of an online social network. *Journal of marketing research*, 48(3):425–443, 2011.

[114] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[115] Michael Kearns, Michael Littman, and Satinder Singh. Graphical models for game theory. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, 2001.

[116] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[117] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[118] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[119] Jon Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. 2007.

[120] Jan Kmenta. Mostly harmless econometrics: An empiricist's companion, 2010.

[121] D. Koller and N. Friedman. *Probabilistic graphical models: Principles and techniques*. MIT Press, 2009.

[122] Dirk Koschützki, Katharina Anna Lehmann, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl, and Oliver Zlotowski. Centrality indices. In *Network analysis*, pages 16–61. Springer, 2005.

[123] Gueorgi Kossinets and Duncan J Watts. Origins of homophily in an evolving social network. *American journal of sociology*, 115(2):405–450, 2009.

[124] Balázs et al. Kovács. The paradox of publicity: How awards can negatively affect the evaluation of quality. volume 59, pages 1–33. SAGE Publications Sage CA: Los Angeles, CA, 2014.

[125] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, page 201320040, 2014.

[126] Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, pages 741–750. ACM, 2009.

[127] Brenden Lake and Joshua Tenenbaum. Discovering structure by learning sparse graph. In *Proceedings of the Annual Cognitive Science Conference*, 2010.

[128] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[129] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, May 2015.

[130] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2001.

[131] Dokyun Lee and Kartik Hosanagar. Impact of recommender systems on sales volume and diversity. 2014.

[132] John Boaz Lee, Ryan A Rossi, Sungchul Kim, Nesreen K Ahmed, and Eunyee Koh. Attention models in graphs: A survey. *arXiv preprint arXiv:1807.07984*, 2018.

[133] Yan Leng, Larry Rudolph, Jinhua Zhao, and Haris N Koutsopolous. Synergistic data-driven travel demand management based on phone records. 2017.

[134] Yan Leng, Dominiquo Santistevan, and Alex Pentland. Familiar strangers: the collective regularity in human behaviors. *arXiv preprint arXiv:1803.08955*, 2018.

[135] Yan Leng, Yehonatan Sella, Rodrigo Ruiz, and Alex Pentland. Contextual centrality: Going beyond network structures. *arXiv preprint arXiv:1805.12204*, 2018.

[136] Kevin Lewis, Marco Gonzalez, and Jason Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1):68–72, 2012.

[137] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

[138] Lusi Li, Jianqing Chen, and Srinivasan Raghunathan. Recommender system rethink: Implications for an electronic marketplace with competing manufacturers. *Information Systems Research*, 29(4):1003–1023, 2018.

[139] Ernest Liu. Industrial policies and economic development. 2017.

[140] Ilan Lobel and Evan Sadler. Information diffusion in networks through social learning. *Theoretical Economics*, 10(3):807–851, 2015.

[141] Ilan Lobel and Evan Sadler. Preferences, homophily, and social learning. *Operations Research*, 64(3):564–584, 2015.

[142] Haokai Lu and James Caverlee. Exploiting geo-spatial preference for personalized expert recommendation. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 67–74. ACM, 2015.

[143] Christopher Lynn and Daniel D Lee. Maximizing influence in an ising network: A mean-field optimal solution. In *Advances in Neural Information Processing Systems*, pages 2495–2503, 2016.

[144] Liye Ma, Ramayya Krishnan, and Alan L Montgomery. Latent homophily or social influence? an empirical analysis of purchase within a social network. *Management Science*, 61(2):454–473, 2014.

[145] Agustino Martınez-Antonio and Julio Collado-Vides. Identifying global regulators in transcriptional regulatory networks in bacteria. *Current opinion in microbiology*, 6(5):482–489, 2003.

[146] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro. Connecting the dots: Identifying network structure via graph signal processing. *IEEE Signal Processing Magazine*, 36(3):16–43, 2019.

[147] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[148] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *CoRR*, abs/1406.6247, 2014.

[149] Markus Mobius and Tanya Rosenblat. Social learning in economics. *Annu. Rev. Econ.*, 6(1):827–847, 2014.

[150] Wendy W Moe and Michael Trusov. The value of social dynamics in online product ratings forums. *Journal of Marketing Research*, 48(3):444–456, 2011.

[151] Andrea Montanari and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010.

[152] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model CNNs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[153] F. Monti, M. M. Bronstein, and X. Bresson. Geometric matrix completion with recurrent multi-graph neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[154] Seth A Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41. ACM, 2012.

[155] Mark EJ Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[156] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[157] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3):245–251, 2010.

[158] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst. Graph signal processing: Overview, challenges and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

[159] Wei Pan, Nadav Aharony, and Alex Pentland. Composite social network for predicting mobile apps installation. 2011.

[160] Wei Pan, Yaniv Altshuler, and Alex Pentland. Decoding social influence and the wisdom of the crowd in financial trading network. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 203–209. IEEE, 2012.

[161] Alex Pentland. *Social physics: How good ideas spread-the lessons from a new science*. Penguin, 2014.

[162] Paolo Pin and Brian W Rogers. Stochastic network formation and homophily. 2016.

[163] Maoying Qiao, Jun Yu, Wei Bian, Qiang Li, and Dacheng Tao. Adapting stochastic block models to power-law degree distributions. *IEEE transactions on cybernetics*, 49(2):626–637, 2018.

[164] Liangfei Qiu, Zhan Shi, and Andrew B Whinston. Learning from your friends' check-ins: An empirical study of location-based social networks. *Information Systems Research*, 29(4):1044–1061, 2018.

[165] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.

[166] Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American s. tistical Association*, 79(387):516–524, 1984.

[167] Ryan A Rossi and Nesreen K Ahmed. Role discovery in networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):1112–1131, 2015.

[168] Roland T Rust and Ming-Hui Huang. The service revolution and the transformation of marketing science. *Marketing Science*, 33(2):206–221, 2014.

[169] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

[170] Paul A Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71, 1938.

[171] Moshen Shahriari and Mahdi Jalili. Ranking nodes in signed social networks. *Social Network Analysis and Mining*, 4(1):172, 2014.

[172] Cosma Rohilla Shalizi and Andrew C Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research*, 40(2):211–239, 2011.

[173] Chao Shang, Qinqing Liu, Ko-Shin Chen, Jiangwen Sun, Jin Lu, Jinfeng Yi, and Jinbo Bi. Edge attention-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1802.04944*, 2018.

[174] Chuan Shi, Binbin Hu, Wayne Xin Zhao, and S Yu Philip. Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 31(2):357–370, 2019.

[175] Zhan Shi and Andrew B Whinston. Network structure and observational learning: Evidence from a location-based social network. *Journal of Management Information Systems*, 30(2):185–212, 2013.

[176] D. I Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, May 2013.

[177] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[178] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009, 2009.

[179] Kar Yan Tam and Shuk Ying Ho. Understanding the impact of web personalization on user information processing and decision outcomes. *MIS quarterly*, pages 865–890, 2006.

[180] Jiliang Tang, Xia Hu, and Huan Liu. Social recommendation: a review. *Social Network Analysis and Mining*, 3(4):1113–1133, 2013.

[181] **Leng Yan**, Xiaowen Dong, Esteban Moro, and Alex Pentland. The rippling effect of social influence via phone communication network. 2018.

[182] **Leng Yan**, Xiaowen Dong, and Alex Pentland. Learning quadratic games on networks. *arXiv preprint arXiv:1811.08790. Under review in Nature Communications*, 2019.

[183] **Leng Yan**, Haris Koutsopoulos, and Jinhua Zhao. Profiling presence patterns and segmenting user locations from cell phone data. *arXiv preprint arXiv:1805.12208*, 2018.

[184] **Leng Yan**, Alejandro Noriega, Alex Pentland, Ira Winder, Nina Lutz, and Luis Alonso. Analysis of tourism dynamics and special events through mobile phone metadata. *Data for good exchange*, 2016.

[185] **Leng Yan**, Larry Rudolph, Alex'Sandy' Pentland, Jinhua Zhao, and Haris N Koutsopolous. Managing travel demand: Location recommendation for system efficiency based on mobile phone data. *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, 2017.

[186] **Leng Yan**, Martin Saveski, Dean Eckles, and Alex Pentland. Observational causal inference with network information. *NeurIPS 2019, Graph Representation Learning Workshop.*, 2019.

[187] **Leng Yan**, Tara Sowriraja, and Alex Pentland. Measuring social influence within and across multi-dimensional homophilous communities. *Presented at International Conference for Computational Social Science (IC2S2)*, 2019.

[188] Jeffrey Travers and Stanley Milgram. The small world problem. 1967.

[189] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, page 201116502, 2012.

[190] R. van den Berg, T. N. Kipf, and M. Welling. Graph convolutional matrix completion. In *arXiv:1706.02263*, 2017.

[191] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[192] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[193] U. von Luxburg. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395–416, Dec 2007.

[194] H.-T. Wai, A. Scaglione, and A. Leshem. Active sensing of social networks. *IEEE Transactions on Signal and Information Processing over Networks*, 2(3):406–419, Sep 2016.

[195] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[196] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[197] Zheng Wen, Branislav Kveton, Michal Valko, and Sharan Vaswani. Online influence maximization under independent cascade model with semi-bandit feedback. In *Advances in Neural Information Processing Systems*, pages 3026–3036, 2017.

[198] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3:2522, 2013.

[199] Matthew J Williams and Mirco Musolesi. Spatio-temporal networks: reachability, centrality and robustness. *Royal Society open science*, 3(6):160196, 2016.

[200] Stanley Wong. *Foundations of Paul Samuelson's Revealed Preference Theory: A study by the method of rational reconstruction.* Routledge, 2006.

[201] Bowei Yan and Purnamrita Sarkar. Covariate regularized community detection in sparse graphs. *Journal of the American Statistical Association*, (just-accepted):1–29, 2019.

[202] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 599–608. IEEE, 2010.

[203] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 325–334. ACM, 2011.

[204] H Peyton Young. Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American economic review*, 99(5):1899–1924, 2009.

[205] Donghan Yu, Yong Li, Fengli Xu, Pengyu Zhang, and Vassilis Kostakos. Smartphone app usage prediction using points of interest. *arXiv preprint arXiv:1711.09337*, 2017.

[206] Xiao Yu, Xiang Ren, Yizhou Sun, Quanquan Gu, Bradley Sturt, Urvashi Khandelwal, Brandon Norick, and Jiawei Han. Personalized entity recommendation: A heterogeneous information network approach. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 283–292. ACM, 2014.

[207] Juanjuan Zhang. The sound of silence: Observational learning in the us kidney market. *Marketing Science*, 29(2):315–335, 2010.

[208] Juanjuan Zhang and Peng Liu. Rational herding in microloan markets. *Management science*, 58(5):892–912, 2012.

[209] Feng Zhu and Xiaoquan Zhang. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, 74(2):133–148, 2010.

[210] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67, Part 2:301–320, 2005.

Appendix A

Interpretable Stochastic Block Influence Model: measuring social influence among homophilous communities

Abstract

Decision-making on networks can be explained by both homophily and social influence. While homophily drives the formation of communities with similar characteristics, social influence occurs both within and between communities. Social influence can be reasoned through role theory, which indicates that the influences among individuals depend on their roles and the behavior of interest. To operationalize these social science theories, we empirically identify the homophilous communities and use the community structures to capture the "roles", which affect the particular decision-making processes. We propose a generative model named Stochastic Block Influence Model and jointly analyze both the network formation and the behavioral influence within and between different empirically-identified communities. To evaluate the performance and demonstrate the interpretability of our method, we study the adoption decisions of microfinance in an Indian village. We show that although individuals tend to form links within communities, there are strong positive and negative social influences between communities, supporting the weak tie theory. Moreover, we find that communities with shared characteristics are associated with positive influence. In contrast, the communities with a lack of overlap are associated with negative influence. Our framework facilitates the quantification of the influences underlying decision communities and is thus a useful tool for driving information diffusion, viral

marketing, and technology adoptions.[1] Social influence; Homophily; Stochastic Block Model; Community structure; Generative model

A.1 Introduction

We are living in an increasingly connected society [188, 18, 134**?**]. The connections among individuals foster information diffusion and enable the inter-dependencies in decision-making among peers. Therefore, understanding and modeling how hidden social influence changes individuals' decision-making are essential and critical for many practical applications, such as viral marketing, political campaigns, and large-scale health behavioral change [79, 160**?** , 135].

Homophily, the tendency of similar individuals to associate together, widely exhibits in various types of social networks, and governs the outcomes of many critical network-based phenomena [147, 123, 69]. Salient features for homophily come from a wide range of sources, including age, race, social class, occupational, and gender [147]. The complex nature of social relationships and high-dimensional characteristics of individuals thus determine the multi-dimensionality of homophily [41]. Homophily results in locally clustered communities and may affect network dynamics, such as information diffusion and product adoption. The Block Model has been applied to low-dimensional, pre-defined homophilous features and provides a building block to uncover underlying community structures[2] with high-dimensional homophily empirically [1].

Social influence is widely studied in economics and computer science literature due to its importance in understanding human behavior. In economics, researchers focus on causally disentangling social influence from homophily with randomization strategies, such as propensity score matching [14], behavioral matching [**?**] and regression adjustment [12]. In the computer science literature, researchers focus on maximizing the likelihood of the diffusion path of influence by proposing different generative processes [91, 92, 154, 202]. These works focus on the strength or the pathways of social influence, and they do not link social influence to the underlying

[1] This work is joint with Tara Sowrirajan and Alex Pentland.

[2] In this appendix, we use community and block interchangeably.

homophilous communities and the network formation process.

There exist two theories explaining how local communities affect information diffusion [198] and contagion in decision-making [**?** **?**]. On the one hand, homophily and the requirement of social reinforcement for behavioral adoption in complex contagion theory indicate that influence tends to be localized in homophilous communities [147, 59]. In other words, behavioral diffusion and network formation are endogenous, explaining the phenomenon of within-community spreading [162, 198]. On the other hand, the weak ties theory [96] implies that bridging ties between communities facilitate the spreading of novel ideas. As empirical evidence, Ugander shows that reinforcement from the multiple communities, rather than from the same communities, predicts higher adoption rates [189]. With these two competing theories, we seek to understand whether social influence spreads locally within each homophilous community or globally to other communities taking advantage of the long ties.

Role theory posits that "the division of labor in society takes the form of interaction among heterogeneous specialized positions" [36]. That is to say, depending on the social roles and the behavior of interest, the underlying interactions and norms for decision-making are different. Motivated by this proposition, we aim to develop a method to associate social influence with the underlying communities, which are associated with the behavior of interest. To formalize this idea, we propose a generative model to understand how social influence impacts decision-making by inferring the spreading of influence across empirically-identified blocks. Our framework jointly uncovers the underlying blocks and infers two types of relationships across these blocks: social interaction and social influence. Different from the Stochastic Block Model, the observed individual decisions are used to inform the communities, as complementary to the observed network. Along with this, we infer an influence matrix as the social influence across different communities. This influence matrix reveals the hidden social influence at the community level, which would otherwise be impossible to observe and generalize.

As a case study, we experiment on the diffusion of microfinance in an Indian village and perform extensive analysis on the influence matrix estimated from the

model. We find that even though social relationships are denser within communities, social influence mainly spreads across communities. This may be explained by the importance of cross-community weak ties [96] and the strength of structural diversity [189]. Our generative framework and subsequent understanding of how social influence operates are informative for practical applications, such as viral marketing, political campaigns, and large-scale health-related behavioral change [79, 160**?**].

Contributions To summarize, the Stochastic Block Influence Model (SBIM) developed in our study makes the following contributions to the literature:

- SBIM integrates networks, individual decisions, and characteristics into the generative process. It jointly infers two types of relationships among empirically-identified communities: social connection and social influence. Moreover, our model flexibly accommodates both positive and negative social influences.

- Our model is motivated by role theory, which posits that individuals make decisions depending on the context of the decision type [36], e.g., adopting microfinance as opposed to adopting healthy habits. To achieve this, we allow the underlying community to vary with the behavior of interest.

- We perform a case study on the adoption of microfinance in an Indian village. Moreover, we demonstrate the interpretability of our model with a detailed analysis of the influence structure.

- The analysis from our study can be used for designing network interventions and marketing strategies. For example, we show that communities with smaller overlaps in characteristics exert negative influences on one another. Therefore, marketing firms should encourage individuals to communicate with neighbors in the same community, such as inviting these individuals together to an informational event to promote the positive influence among them.

- SBIM bridges the rich Stochastic Block Model and the social contagion literature.

It opens up future opportunities to adapt to other variations of SBM, e.g., degree-regularized SBM [83] or SBM adjusted for power-law distributions [163].

The remaining sections are organized as follows. We describe the literature in Section A.2. In Section A.3, we introduce the proposed Stochastic Block Influence Model. Then, we test the method in Section A.4 and analyze the results on a real-world data set in Section A.4. In Section A.5, we summarize this appendix with practical applications and future work.

A.2 Related literature

Contagion models There are two prominent theories in the literature for explaining the propagation of social influence [189, 44, 14**?**], i.e., simple contagion and complex contagion. Simple contagion theory assumes that individuals will adopt the behavior as long as they have been exposed to the information [96], which is a sensible model for epidemics and information spreading. Complex contagion theory, on the other hand, requires social reinforcement from neighbors to trigger the adoption [59]. Many studies have shown that complex contagion explains behaviors such as registration for health forums [58].

These exposure-based models bear analytical simplicity, however, do not allow social influence to be negative, i.e., the adoption decision of one's neighbors might decrease, rather than increase, the likelihood of one's adoption decision. Moreover, they typically are not able to capture the heterogeneity of social influence [135]. In this chapter, we propose a model to account for negative and heterogeneous influence.

Stochastic Block Model The Stochastic Block Model is a statistical model for studying latent cluster structures in network data [1]. SBM generalizes the Erdos-Renyi random graph model with higher intra-cluster and lower inter-cluster probability. The traditional SBM only infers the community structures from network connections. However, when contextual information on nodes is available, leveraging information from different sources facilitates the inference. In recent statistics literature, there

has been some interesting work on utilizing covariates to infer the block structures. For example, Binkiewicz et al. present a covariate-regularized community detection method to find highly connected communities with relatively homogeneous covariates [37]. They balance the two objectives (i.e., the node covariance matrix and the regularized graph laplacian) with tuned hyper-parameters. Yan et al. propose a penalized optimization framework by adding a k-means type regularization [201]. This framework enforces that the estimated communities are consistent with the latent membership in the covariate space.

Though these variations to SBM utilize auxiliary information on individual nodes, they specify the importance of recovering the network and the smoothness of covariates on the network, on an ad-hoc basis. Different from these models, we take advantage of role theory [36] and utilize the decision-making process on the network that could also inform community detection. For example, let us assume professional communities are more useful for the adoption of technologies at work, and social communities are more useful for the adoption of social apps. The underlying communities depend on the role and behavior of interest because social influence spreads through some specific network links in different applications.

A.3 Methodology

A.3.1 Stochastic Block Influence Model

Notations Assume a random graph $G(\mathcal{V}, \mathcal{E})$ with N individuals in node set $\mathcal{V}$ and edge set $\mathcal{E}$. It is partitioned into C disjoint blocks $(\mathcal{V}_1, ..., \mathcal{V}_C)$, and the proportion of nodes in each block c is π_c, and $\sum_{c=1}^{C} \pi_c = 1$. $\mathbf{A} \in \mathbb{R}^{N \times N}$ represents the adjacency matrix. $\mathbf{A}_{ij} = 1$ if i and j are connected, and $\mathbf{A}_{ij} = 0$ otherwise. Let matrix $\mathbf{B} \in \mathbb{R}^{C \times C}$ denote the inter-block and intra-block connection probability matrix. Let $\mathbf{M}_i$ be the block assignment of individual i and summing over C blocks, we have $\sum_{k=1}^{C} \mathbf{M}_{ik} = 1$. Together, we combine the block vector of all individuals in the matrix $\mathbf{M} \in \mathbb{R}^{N \times C}$. Therefore, the probability of a link between v_i and v_j between two separate blocks

$\mathcal{V}_k$ and $\mathcal{V}_l$ as $P\big((v_i, v_j) \in \mathcal{E} | v_i \in \mathcal{V}_k, v_j \in \mathcal{V}_l\big) = p_{ij}$. $\mathbf{y} \in \mathbb{R}^N$ is a binary vector representing individuals' adoption behaviors. Let $\mathbf{X}_i \in \mathbb{R}^D$ represent demographic features, where D is the number of covariates. We use $\mathbf{F} \in \mathbb{R}^{C \times C}$ to represent the block-to-block influence matrix. Finally, $\mathbf{h}$ is a binary vector, capturing whether or not each individual is aware of the product at the beginning of the observational period. For a new product, $\mathbf{h}$ is sparse, while for a mature product, $\mathbf{h}$ is dense.

Model formulation Extending SBM to utilize the network, adoption decisions, and sociodemographic features, we propose the Stochastic Block Influence Model, abbreviated as SBIM. Linking the latent communities to their sociodemographic composition, we reveal the underlying nature of high-dimensional homophily in a data-driven fashion rather than using pre-defined communities using observed sociodemographics, e.g., race or occupation. Solely using pre-defined homophilous characteristics does not aptly capture the multiplex characteristics that define individuals and their social ties. In other words, individuals are associated with different communities, each of which is formed by various homophilous characteristics. Neighbors belonging to different communities may influence the focal individuals differently.

Let us illustrate this using the adoption of microfinance in an Indian village. It is reasonable to posit that several traits define the diverse nature of individuals - different professions, castes, education levels, and a variety of other demographic features. Let us take one particular individual, who is an educated worker of a lower caste, for example. This individual belongs with varying degrees of affiliation to different communities: perhaps most strongly affiliated to a group of a certain level of education and less strongly affiliated with another group of a majority of a lower caste. This mixed membership captures the realistic nature of our social relationships and characteristics. Within such a village with multi-dimensional homophily, how can we understand who influences this individual and what processes are involved in that individual's decision making? Specifically, she could be influenced both by neighbors belonging to different communities characterized by specific educational backgrounds, professions, and castes. The data-driven multi-dimensional block aspect of the model

allows us to capture these critical, hidden relationships.

Next, we formalize our model. To jointly infer how influence spreads within and across communities, we desire a model with the following properties:

1. The model leverages both the observed friendship network structure and the adoption behavior to infer the underlying communities.

2. The link formation and social influence between two individuals are jointly determined by their underlying communities.

For each individual pair $\{i, j\}$, depending on their community assignment vectors, the predicted link $\tilde{\mathbf{A}}_{ij}$ is generated according to the connection probability matrix, $\mathbf{B}$. In particular, the probability of the existence of a link between i and j is,

$$\mathbb{P}(\tilde{\mathbf{A}}_{ij} = 1|\mathbf{M}, \mathbf{B}) = (\mathbf{MBM}^T)_{ij}. \tag{A.1}$$

Next, we discuss how our model incorporates individual characteristics and adoption decisions. The adoption likelihood depends on individuals' characteristics and on the influence of their neighbors who have already adopted [113]. The generative model builds upon the communities a particular individual i, and i's neighbors belong to, as well as the community-to-community matrix $\mathbf{F}_{ij}$. Each individual makes a decision on whether or not to adopt in order to maximize her utility. The utility of i depends on her own preferences and the aggregated influence from neighbors. The pairwise influence depends on the communities i and her neighbors belong to. We illustrate how influence and communities affect one's decision-making in Figure A-1. Let us consider individual A, who has three friends, B, C, and D, belonging to a lower socioeconomic status (SES) group (as colored in red), and one friend, E, belonging to a higher SES group (as colored in blue). The adoption likelihood of A is a function of her own preferences as well as the influence from her friends B, C, D, and E. The strength of the influence depends on the corresponding communities of A and her friends (B, C, D, and E).

More generally, the adoption likelihood of a user, $\hat{\mathbf{y}}$, is defined as,

$$\hat{\mathbf{y}}_i = \text{logit}\big(\boldsymbol{\beta}\mathbf{X}_i + \sum_j \big((\mathbf{M}\mathbf{F}\mathbf{M}^T) \circ ((\mathbf{h} \otimes \mathbf{1}) \circ \mathbf{A})\big)_{ji} + \epsilon_i\big), \tag{A.2}$$

where $\circ$ is the element-wise matrix multiplication. The first term, $\boldsymbol{\beta}\mathbf{X}_i$, measures the adoption decision conditioned on i's sociodemographic features if there were no social influence, where $\boldsymbol{\beta} \in \mathbb{R}^D$ and D is the dimension of the covariates. The second term aggregates the influence of i's neighbors. ϵ_i is the idiosyncratic error term. Without loss of generality, we assume $\epsilon_i \sim \mathcal{N}(0, 1)$.

For a mature product that everyone is aware of, we can simplify Equation (A.2) as,

$$\hat{\mathbf{y}}_i = \text{logit}\Big(\boldsymbol{\beta}\mathbf{X}_i + \sum_{j=1}^{N} \big((\mathbf{M}\mathbf{F}\mathbf{M}^T) \circ \mathbf{A}\big)_{ji} + \epsilon_i\Big). \tag{A.3}$$

Equation (A.2) only accounts for the influence among direct neighbors. Note that in a small-scale network, it is reasonable to assume that there does not exist higher-order social influence. In a large-scale network, Leng et al. show that social influence spreads beyond immediate neighbors [**?**]. For these applications, our model can be easily adapted to higher-order influence by summing up the powers of the adjacency matrix A to account for multiple degrees of separation [135].

A.3.2 Generative process

For the full network, the model assumes the following generative process, which defines a joint probability distribution over N individuals, based on node-wise membership matrix $\mathbf{M}$, block-to-block interaction matrix $\mathbf{B}$, block-to-block influence matrix $\mathbf{F}$, attributes' coefficients $\boldsymbol{\beta}$, observed friendship network $\mathbf{A}$, observed attributes $\mathbf{X}$, observed adoption decision $\mathbf{y}$.

1. For each node $v_i \in \mathcal{V}$, draw a C-dimensional mixed membership vector $\mathbf{M}_i \sim$ Dirichlet(c).

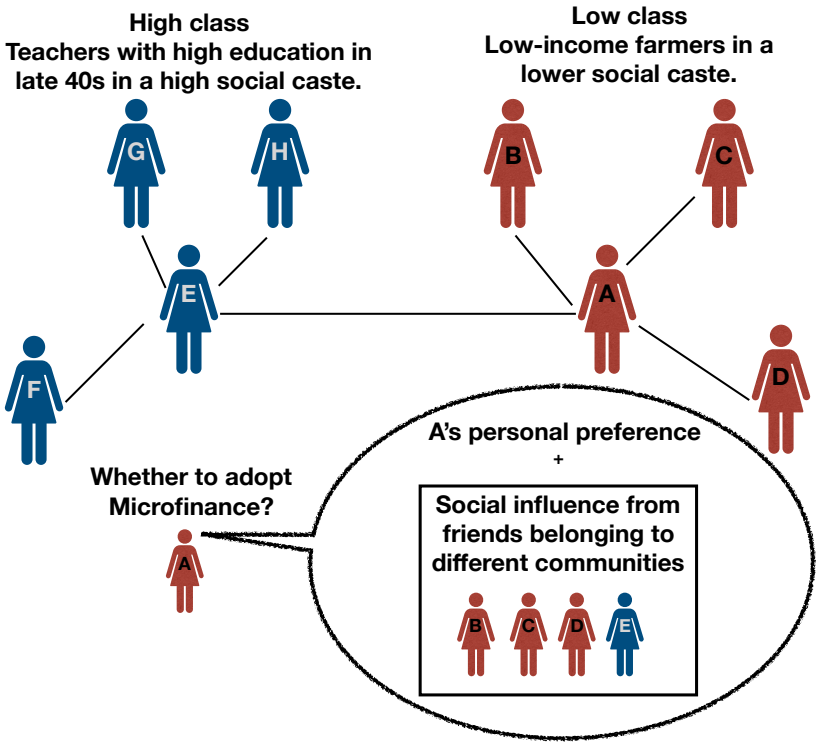


Figure A-1: **Graphical representation of the Stochastic Block Influence Model (SBIM).** Assume there are two communities, a high socioeconomic status (SES) group (colored in dark blue) and low SES group (colored in dark red), characterized by multi-dimensional sociodemographic features. The two groups have higher intra-class connection probability and lower inter-class connection probabilities. The decision-making of A is jointly influenced by her preferences, as well as her neighbors from the same and different communities.

2. For the connection probability from community k to l in the block-to-block connectivity matrix, draw $\mathbf{B}_{kl} \sim \text{Beta}(a, b)$.

3. For the influence from community k to l in the block-to-block influence matrix, draw $\mathbf{F}_{kl} \sim \mathcal{N}(\mu_F, \sigma_F)$.

4. For each attribute in $\boldsymbol{\beta}$ indexed by d, draw the coefficient $\boldsymbol{\beta}_d \sim \mathcal{N}(\mu_b, \sigma_b)$.

5. Draw the connection between each pair of nodes v_i and v_j, $\hat{\mathbf{A}}_{ij}$, according to Equation (A.1).

6. Draw the adoption decision $\hat{\mathbf{y}}_i$, according to Equation (A.2).

For abbreviation, we denote $\mathcal{Z}$ as set of the hidden variables, $\mathcal{Z} = \{\mathbf{M}, \boldsymbol{\beta}, \mathbf{B}, \mathbf{F}\}$ and θ as the set of hyperparameters, where $\theta = \{c, a, b, \mu, \sigma, \mu_b, \sigma_b\}$.

The posterior distribution defined by the generative model is a conditional distribution of the hidden block structure and relationships given the observed friendship network and adoption behavior, which decomposes the agents into C overlapping blocks. The posterior will place a higher probability on configurations of the community membership that describe densely connected communities as well as stronger (positive or negative) influences. We present a visualization in Figure A-3, which illustrates that the posterior superimposes a block structure on the original network. The details of the data we use are described in Section A.4.

Inference The posterior of SBIM is intractable, similar to many hierarchical Bayesian models [31]. Therefore, we use the Markov Chain Monte Carlo (MCMC) algorithm as an approximate statistical inference method to estimate the parameters. MCMC draws correlated samples that converge in distribution to the target distribution and are generally asymptotically unbiased.

There are different MCMC methods, including Gibbs sampling, Metropolis-Hastings, Hamiltonian Monte Carlo, and No-U-Turn Sampler (NUTS). Gibbs sampling and Metropolis-Hastings methods converge slowly to the target distribution as they explore the parameter space by random walk [102]. HMC suppresses the random walk

behaviors with an auxiliary variable that transforms the problem by sampling to a target distribution into simulating Hamiltonian dynamics. However, HMC requires the gradient of the log-posterior, which has a complicated structure in our model. Moreover, it requires a reasonable specification of the step size and a number of steps, which would otherwise result in a substantial drop in efficiency [103].

Therefore, we apply NUTS, a variant to the HMC method, to eliminate the need for choosing the number of steps by automatically adapting the step size. Specifically, NUTS builds a set of candidate points that spans the target distribution recursively and automatically stops when it starts to double back and retrace its steps [103]. We use the NUTS algorithm implemented in Python PyMC3 [169].

A.4 Experiments

Data description We study the adoption of microfinance in an Indian village collected by the Abdul Latif Jameel Poverty Action Lab (J-PAL) [23][3]. In 2007, a microfinance institution introduced a microfinance program to some selected Indian villages. In early 2011, they collected information about whether or not the villagers had adopted microfinance. Because the village is fairly small (257 villagers) and microfinance had been on the market for four years when JPAL collected individuals' adoption decisions, it is reasonable to assume that everyone in the village was aware of microfinance, which is hence a mature product. Therefore, we use Equation (A.3) as the decision-making function. The data contains information about self-reported relationships among households and other amenities, including village size, quality of access to electricity, quality of latrines, number of beds, number of rooms, the number of beds per capita, and the number of rooms per capita. These types of demographic features are used as the independent variables. The outcome variable is the adoption decision of = microfinance. The microfinance institution asked the villagers to self-report other villagers they considered as friends.

[3]The village we study is indexed by 64.

Baseline We use the Random Forest with sociodemographics and the hidden community learned by spectral clustering on the adjacency matrix as the independent variables. In this way, we use the same information in SBIM and the baseline. Spectral clustering uses the second smallest eigenvector of the graph laplacian as the semi-optimal partition [156].

Model training To train our model and evaluate the performance for a particular C, the number of block, we cross-validated by randomly splitting the data into 75% training samples and 25% test samples. We repeat this process ten times. With NUTS, we obtain the point estimates for all latent variables in $\mathcal{Z}$[4]. We then re-run our model (as previously described) with all latent variables fixed to the estimates on the test dataset. This step returns the predicted adoption probability for each villager in the test data.

To choose the optimal number of block, we first tune the model for $C \in \{2, 6, 10, 14\}$ and then calculate the average loss. We observe a negative parabolic trend with the loss peaking at its lowest at $C = 10$ blocks, so we use this optimal number of block for further evaluation.

Model evaluation Since the dependent variable in our data is imbalanced, we evaluate our method using the AUC, which is the area under the Receiver-Operating-Characteristics curve plotted by the false positive rate and correct positive rate for different thresholds. We define a loss metric during the training period to select the best configurations. It is formulated by the negative of the standard improvement measure, which is the absolute improvement in performance normalized by the room for improvement. This measure captures the improvement of our method compared to the baseline. Since we have a small test set, a randomly-drawn test set may be harder to predict than others. Measuring the relative improvement ensures that the

[4]Some critical hyperparameters for NUTS are the number of burn-in samples, the number of samples after burn-in, the target acceptance probability, and the number of chains. For all of our NUTS sampling runs, we burn 3,000 samples to ensure that MCMC mostly converges to the actual posterior distribution. The number of samples after burn-in is 500; usually, only less than ten samples (among the 500) are diverging. Next, we select the target acceptance probability to be 0.8. At the end of each run, we average across the 500 samples to derive point estimates for all latent variables.

composition of the test set does not bias the performance due to sample variation. This metric is formulated by

$$L = \frac{\text{Baseline}_{\text{ test AUC}} - \text{SBIM}_{\text{ test AUC}}}{1 - \text{Baseline}_{\text{ test AUC}}}, \tag{A.4}$$

where the AUC of the baseline and SBIM on the test split in cross-validation are represented as $\text{Baseline}_{\text{ test AUC}}$ and $\text{SBIM}_{\text{ test AUC}}$, respectively.

Our model has seven hyperparameters, $\theta = \{c, a, b, \mu, \sigma, \mu_b, \sigma_b\}$[5]. Since the parameter space is large, we adapt a bandit-based approach to tune the parameters developed called Hyperband [137]. The Hyperband algorithm adaptively searches for configurations and speeds up the process by adaptive resource allocation and early-stopping. Our adaptation of this algorithm allows each configuration tested to run with full resources due to the sampling procedure used in our methodology, allowing NUTS to run consistently across all configurations.

Performance We compare the performance of our model with the baseline in Table A.1. We observe that our method outperforms random forest in the test set by 13.8% by the improvement metric in Equation (A.4). Both models overfit the training set and the baseline overfit comparatively more.

Table A.1: Model and baseline performance

	Mean	Standard deviation
Baseline train AUC	0.901	0.010
SBIM train AUC	0.805	0.022
Baseline test AUC	0.610	0.095
SBIM test AUC	0.664	0.062

[5]The ranges from which these hyperparameters were sampled are as follows: $\mu_b \in [-2, 2]$, $\sigma_b \in [-0.1, 1]$, $c \in [0.5, 1.5]$, $\mu_F \in [-6, 6]$, and $\sigma_F \in [0.1, 3]$. We let $a, b = 2$ for a reasonable and non-skewed prior.

Analysis and discussions

Size of communities and interaction matrix We present the size of each social block in Figure A-2. Social block two is larger than the other blocks, and the sizes of the rest are similar. This aligns with our intuition that many individuals belong to a majority group while several niches, minority communities also exist. We represent the adjacency matrix sorted by this inferred block index from smallest to largest block in Figure A-3. We see that there are many links within all of the blocks along the diagonal, demonstrating that the block model is meaningful and captures more links within than across blocks. The largest block, furthest along the diagonal, is comparatively sparser.

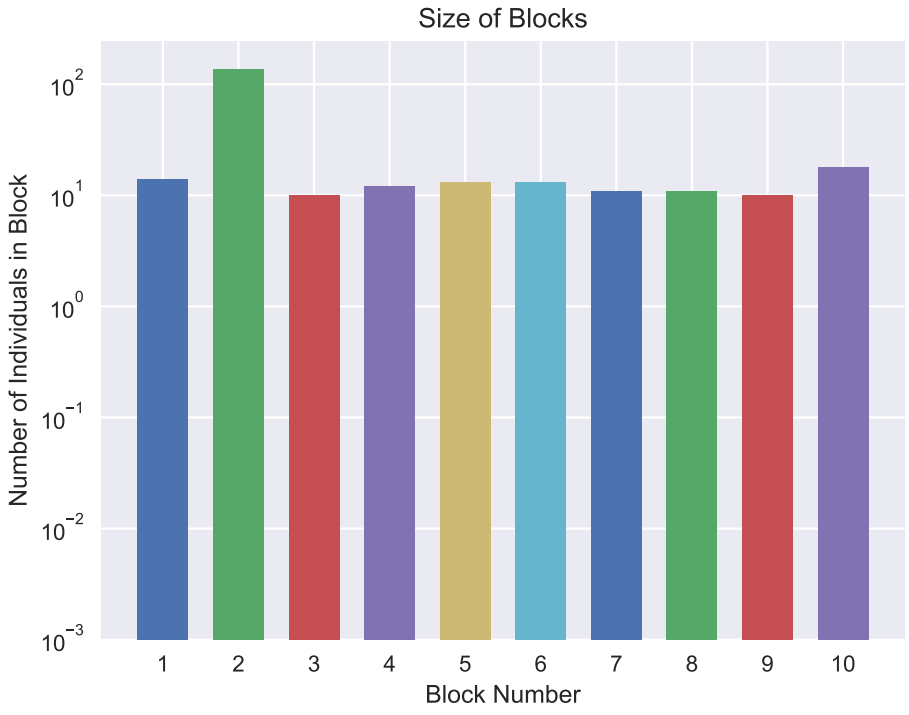


Figure A-2: **Size of each social block.** The y-axis corresponds to the number of individuals in the block, and the x-axis is the corresponding block index.

Block type We can associate individuals' sociodemographic characteristics with the individuals who belong to each block to generalize block type as consisting of

characteristics such as high or low SES, homogeneous or diverse, and skilled or less educated, as depicted in Table A.2. In this example, each block is associated with a qualitative type, and the attributes within that block leading to such characterizations are described. Lower or higher SES blocks are designated by caste composition, education levels, and profession types. Homogeneous or diverse blocks are designated by some professional composition, caste types, mother tongue language composition, gender imbalance, and what fraction of village inhabitants are natives.

We also use diversity and gender ratio to evaluate block characteristics for a specific example in Table A.2 and Figure A-5, in addition to being used to evaluate the group attributes that are associated with different types of influence in Table A.3. More analysis in Figure A-5 is covered in the following section.

We use normalized entropy to measure the diversity of different attributes. Normalized entropy is a metric used to capture the number of types of characteristics within each category while accounting for the frequency of each entity type within a category. It can be formulated by, $Q = -\frac{\sum_{i=1}^{q} p_i \log(p_i)}{\sum_{i=1}^{q} \frac{1}{n_i} \log(\frac{1}{n_i})}$, where q refers to the number of types within a category, p_i refers to the probability of each type i, f refers to the number of occurrences of each type n_i.

The gender ratio (R) is measured within a block and is formulated by $R = \frac{r_m}{r_f}$, where r_m and r_f refer to the number of occurrences of males and females respectively. Thus, since R is the ratio of males to females in a block, both a high or low gender ratio correspond to a high gender imbalance.

Influence matrix and attributes The block-to-block influence, sorted by increasing block size, is displayed in Figure A-4a, where the strength of social influence, allowed to be either positive or negative, is shown. We can see some blocks influence other blocks ranging from strong negative influence to no influence, and to strong positive influence.

The total influence into and out of each block is depicted in Figure A-4b, which allows us to evaluate the aggregated influence a block receives and spreads (net positive, negative, or neutral). For example, we can see diverse, low-SES block five and senior,

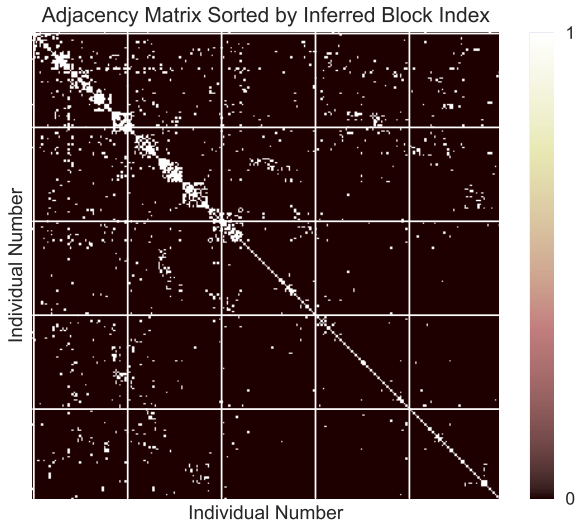


Figure A-3: Adjacency matrix sorted by the inferred block index. The x-axis and y-axis correspond to the indices of individuals. The white and black cells correspond to the existence and the non-existence of edges. We can clearly observe the underlying communities from the network.

low-class block six with high output levels of positive influence, and diverse, middle-SES block eight receives a net high level of negative influence. We observe that some blocks have a stronger outgoing influence than other blocks and can perceive these as positive and negative influence leaders. Similar reasoning applies to characterize blocks that receive a high level of influence as follower blocks, furthermore observing the difference in net incoming and outgoing influence within each block as relating to its role in the block-to-block network. We refer to this to interpret different dynamics between social blocks, in addition to then pairing this information with demographic information to make further evaluations about block characteristics associated with different types of influence.

In Figure A-5, a subset of the sociodemographic features are displayed for each block, where the network of blocks is connected with varying degrees of influence between them. For example, we can see that lower median-age block four negatively influences the older median-age block six. The equal gender ratio block ten positively influences the similarly equal gender ratio block nine. Block ten influences block nine,

where both blocks have similarly high caste diversity. Highly language diverse block six positively influences low language diverse block one. Lower professionally diverse block one negatively influences higher professionally diverse block three.

Table A.2: **Block characteristics example. SES is an abbreviation for socioeconomics status.** The majority refers to the largest subset. Disadvantaged caste refers to lower castes, including the castes OBC (Other Backward Class) and Scheduled. Higher education refers to having education levels at PUC (pre-university course) and having a "degree or above" designation. Moderate and lower education levels include all levels below this, where moderate levels have more SSLC (Secondary School Leaving Certificate) levels, and PUC levels and lower levels have mostly primary school education levels.

Block	Block Type	Attributes
1	Homogeneous, low-SES	only one disadvantaged caste and one language spoken
		low profession diversity and education levels
2	Diverse, skilled, highly-educated	several different castes from many levels
		diverse languages and diverse, high-skilled professions
3	Senior, low-SES	majority disadvantaged caste
		majority low skill-level professions in agriculture
4	Young, low-SES	younger average age, gender imbalanced block
		majority lowest caste members, mostly natives
		higher education
5	Diverse, low-SES	diverse number of disadvantaged castes
		moderate language diversity, moderate education
		majority of jobs in agriculture
6	Senior, low-SES	older average age, diverse in low castes
		two languages spoken, very low education
		lower-skilled professions
7	Homogeneous, low-SES	gender imbalanced, mostly disadvantaged caste
		one language majority
		majority professions in agriculture and sericulture
8	Diverse, middle-SES	mostly one language
		caste diverse but mostly lower castes
		diverse professions
9	Diverse, highly-educated, low-SES	disadvantaged caste majority
		diverse jobs, higher-SES professions (teacher, priest)
		high education level, diverse languages
10	Homogeneous, low-SES	gender-balanced
		majority disadvantaged caste, only one language spoken
		majority of professions in agriculture and sericulture

By analyzing several examples in this manner using block characteristic composition and observing the types and patterns of influence, several general trends arise, as depicted in Table A.3. The block attributes most frequently associated with different types of influence are summarized into key trends. Positive influence occurs when two blocks overlap in the following characteristics: gender distribution, majority castes, professions, high profession diversity, highly educated, highly-skilled jobs, and mother tongue languages. Negative influence frequently occurs when two blocks have a lack of overlap in the following characteristics: gender distribution, caste composition,

Table A.3: **Block attributes associated with different types of influence.** Positive and negative influence refers to the type of influence from one block to another block. Self-influence refers to positive influence within the same block. Overlap refers to overlapping categories, such as caste type, profession type, education levels, or languages spoken.

Attribute	Positive influence	Negative influence	Positive self-influence
Gender	similar gender distribution	gender-imbalanced block is more open to negative influence from gender-balanced block	large gender imbalance
Caste	overlapping majority castes	lack of overlap in caste composition	majority village natives
Profession	profession overlap, in specialty jobs specifically; large professions diversity	professionally diverse block receives negative influence from a less professionally diverse block; lack of professional overlap causes a negative influence	high job diversity and higher-skilled jobs
Education	large overlap in higher education level	higher educated block receives negative influence from less educated block	higher education level
Language	overlapping language	lack of overlap in language	language diversity
Age	none	older-age block can receive negative influence from younger-age block	younger age

profession diversity level, education levels, and average age. Furthermore, the direction of negative influence is most frequently observed from a low-SES block to a high-SES block. Additionally, we frequently observe positive self-influence, which is from a block to itself, and this occurs when a block is characterized by a younger average age, highly-educated, high job diversity, higher-skilled jobs, high language diversity, large gender imbalance, and having a large number of village natives.

These trends, when paired with block type characterizations, lead to interesting associations, such as block-to-block perceptions of lower or higher SES groups with influence. Blocks of the higher SES group designation more frequently received negative influence from lower-SES blocks. Blocks of similar SES, especially higher SES, had a more frequent positive influence between them. High-SES blocks also had more frequent positive self-influence.

These findings suggest some marketing strategies that take into account the underlying communities. For example, the microfinance institution could organize separate information sessions for the high-SES and low-SES groups to take advantage of the positive influence between groups that share similar characteristics, while avoiding the negative influence that occurs across the different communities. Moreover, if the microfinance institution is to introduce the product into other villages (as a

new product), they should send the information to individuals with the following characteristics: (1) high-SES with less low-SES neighbors, (2) individuals who speak a diverse set of languages, and (3) communities with similar gender ratios.

A.5 Applications and future works

Role theory postulates that the interactions of individuals depend on their roles and behaviors of interest. To conceptualize this idea, we use the underlying community structures to capture the "roles", which affect the particular decision-making processes of individuals. Specifically, we develop the Stochastic Block Influence Model, which infers two types of hidden relationships: (1) block-to-block interaction, and (2) block-to-block influence on decision-making. Moreover, our model flexibly allows for both positive and negative social influence. The latter is more common in practice but has been ignored by the contagion models in the literature [59, 116, 23]. In the adoption of microfinance examples we present, the inferred block-to-block influence offers insights into how different social blocks exert influence on individuals' decision-making. The framework has far-reaching practical impacts for understanding patterns of influence across communities and identifying the crucial characteristics of influential individuals for several applications. To name a few:

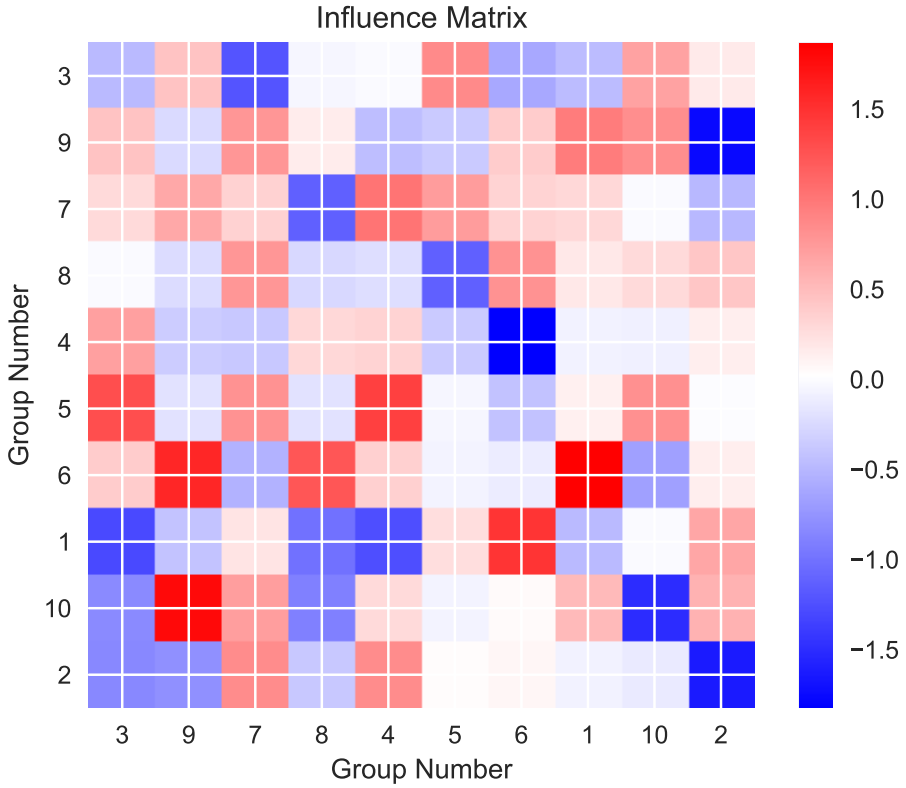
1. Practitioners and researchers can identify the most influential communities (e.g., leaders and followers) and understand the dynamics among different communities that are not available nor observable without our model.

2. Marketing campaigner can investigate in which sociodemographics predict positive or negative social influence, and utilize this information when introducing the product to a new market.

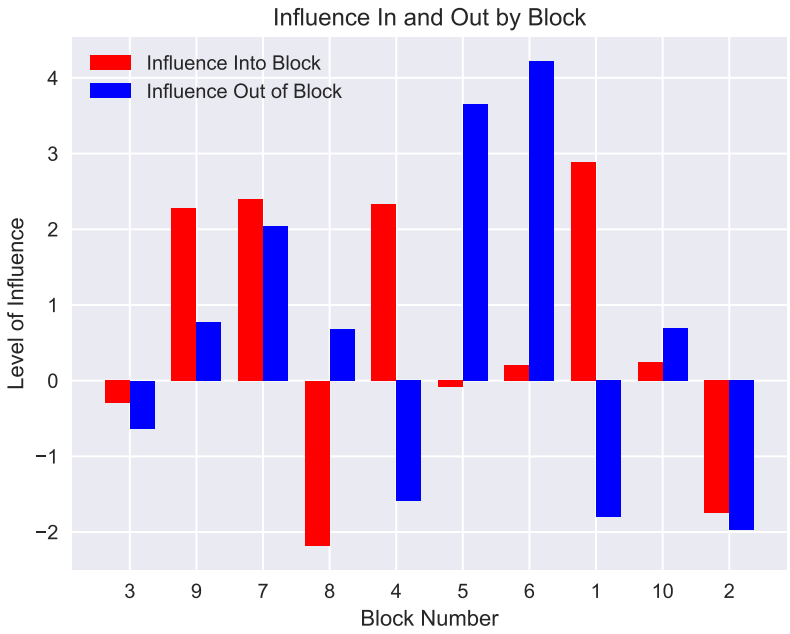
3. Marketing firms can use the influence of each individual to decide whom to target for campaigns [135]. For example, in marketing campaigns, we should advertise to individuals who spread positive aggregate influence.

4. For policy-makers, the behavioral model in this chapter can be used to perform counterfactual predictions for network interventions to predict responses to new policies.

Our method is not without limitations and hence opens up several directions for future studies. First, future research can easily adapt SBIM to accommodate a more complicated stochastic block model, such as a degree-corrected SBM or a power-law regularized SBM. Second, a scalable inference method as an alternative to NUTS sampling will help to improve the efficiency and scalability of SBIM. Third, future research can extend SBIM to a dynamic model, where the influence matrix varies with time and distances from the source of information. Lastly, for computer scientists and social scientists who have access to similar types of data, but in different settings (e.g., different behaviors and collected in different countries), it will be interesting to apply and compare the influence matrices to see if there exists any generalizable pattern to support existing contagion and decision-making theories.



(a) Influence matrix. The x-axis and y-axis correspond to the community index. The values are the strength of influence. The scale from blue to red corresponds with negative to positive influence. Darker colors correspond to a stronger influence. I f the y-axis is $\mathcal{V}_k$ and the x-axis is $\mathcal{V}_l$, the value corresponds to the influence from group $\mathcal{V}_k$ to $\mathcal{V}_l$.



(b) Net influence into and out of each block. The x-axis and y-axis correspond to the block index and the aggregate strength of influence, respectively. The red and blue bars correspond to the in-flow and out-flow of social influence.

Figure A-4: Interaction matrix and influence matrix.

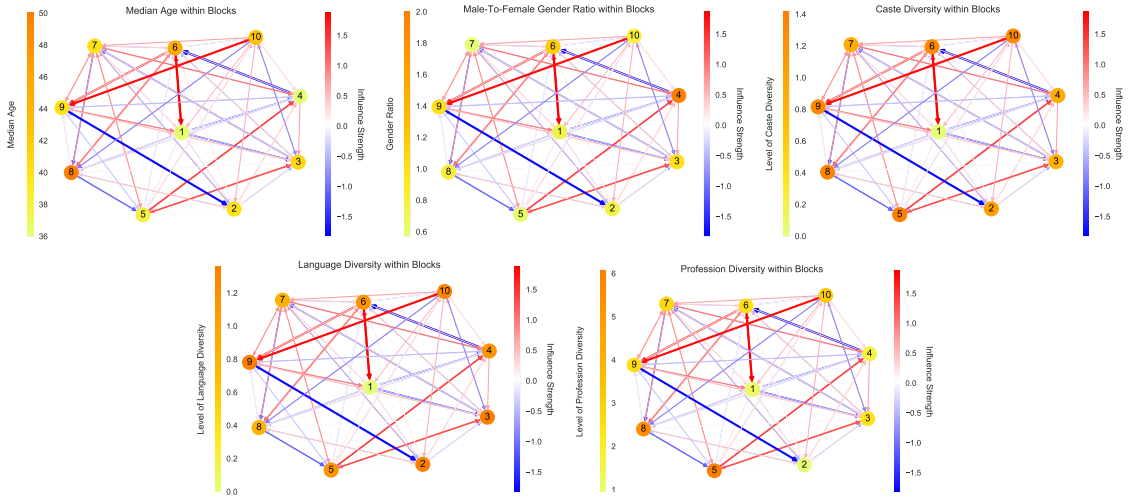


Figure A-5: **Sociodemographic analysis of each social block and the social influence across social blocks.** Each node represents a social block corresponding to the index shown in the previous in Table A.2. The directed links represent the strength of social influence varying from strong negative (blue) to strong positive (red). The color of the node represents a measure of the sociodemographic characteristics within that social block. We display a subset of characteristics, including median age, gender ratio, caste diversity, language diversity, and profession diversity within each block.