Promoting Pro-Social Behavior with End-to-End Data Science

by

Wei Ai

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2020

Doctoral Committee:

Professor Qiaozhu Mei, Chair
Professor Yan Chen
Assistant Professor Daniel Romero
Professor Jieping Ye

Wei Ai

aiwei@umich.edu

ORCID iD: 0000-0001-6271-9430

To my dearest family and friends

ACKNOWLEDGEMENTS

First and foremost, I owe my sincerest gratitude to my advisor, Dr. Qiaozhu Mei, whose "end-to-end" encouragement and guidance accompanied my entire Ph.D. journey. From him, I learned the philosophy of meaningful, insightful, and rigorous research. It is my privilege to have his mentorship and friendship, which will never end.

It is my honor to have Dr. Yan Chen, Dr. Jieping Ye, and Dr. Daniel Romero on my committee, and I would like to thank them for their insights and comments. I am especially indebted to Dr. Yan Chen for her guidance on causal inference and experiment design. I learned a lot from working with her, and she has been a role model for maintaining high standards both in research and in life.

I would also like to thank my undergraduate class advisor Xuanzhe Liu for his continuous support throughout my graduate study, and my colleagues at my alma mater Peking University: Gang Huang, Jian Tang, Yun Ma, Xuan Lu, Huoran Li, Zhenpeng Chen, Shen Shen, and Yanbin Cao. May the fourth be with you.

I am proud of my research family, the Foreseer Research Group at UMSI. I want to express my gratitude to Yang Liu, Danny Tzu-Yu Wu, Zhe Zhao, Xin Rong, Yue Wang, Cheng Li, Sam Carton, Shiyan Yan, Teng Ye, Tera Reynold, Jiaqi Ma, Cristina Garbacea, Zhuofeng Wu, V.G.Vinod Vydiswaran, Xuedong Li, Yutong Xie, Yuhang Zhou, and Jing Xu. It is not a mere list of "cooccurrence." The discussions and debates with them deeply informed my research predilections, and I thank them for sharing the joys and sorrows of my PhD journey.

I enjoyed the interdisciplinary and supportive atmosphere at UMSI, without which this dissertation would not have blossomed. I am especially grateful for Dr. Paul Resnick, whom I greatly admire, for his care and support, and for witnessing every step of my growth. I would also like to thank participants in the IAR (later CS/DSS) seminar and the BEE lab meeting, especially Dr. Kevyn Collins-Thompson, Dr. Ceren Budak, Dr. David Jurgens, Dr. Tanya Rosenblat, and Dr. Alain Cohn for their constructive feedback. I also benefited substantially from my interactions with other members of UMSI, including Chuan-Che (Jeff) Huang, Xuan Zhao, Priyank Chandra, Tamy Guberek, Ming Jiang, Tao Dong, Wenjing (Carrie) Xu, Youyang Hou, Yan Chen, Ark Fangzhou Zhang, Yingzhi Liang, Xinyan Zhao, Yixin Zou, Shiqing (Licia) He, Linfeng Li, Daphne Chang, Iman Yeckehzaare, Chanda Phelan, Katie Cunningham, Tawfiq Ammari, and Roy Chen.

I would like to thank my colleagues during my internships at Kiva, Twitter, and DiDi: Brandon Smith, Harry Pottash, Kate Kleinschmidt, Noah Balmer and Jen Kobayashi at Kiva; Yue Lu, Bingjun Sun, and Lanbo Zhang at Twitter; Lingyu Zhang, Miao Liang, Mei Mei, Tao Song, Lin Zeng, Lulu Zhang, and Chenxi Wang at DiDi. I learned a lot from my industry experience, and this dissertation would not have been possible without these collaborations. I would also acknowledge the contributions of the de-identified users on GitHub, Kiva, and DiDi, without whom my research would be castles in the air. Financial support from NSF and DiDi is gratefully acknowledged.

Finally, I would like to thank my wife Haiyin Liu, and my parents Jianjun Ai and Wenzhen Tian, for their unfailing love and support.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Online platforms provide unprecedented opportunities to nudge pro-social behaviors: They track and host fine-granularity data generated by real-world users, which is a gold mine to understanding and modeling user behaviors. These interactive interfaces and rich functionalities provide excellent flexibility in implementing and delivering the nudge to the users. How shall we unleash the full potential of these platforms to nudge pro-social behaviors?

In this dissertation, we propose an end-to-end data science pipeline that consists of three closely coupled stages: We first analyze user-behaviors with empirical data to discover potential nudges. We then develop recommender systems that maximize the effectiveness of the nudge with personalization. Finally, we implement the nudge in its original context and evaluate the nudge with randomized field experiments. Each stage of the pipeline calls for joint efforts from multiple disciplines, especially causal inference and machine learning. Moreover, the pipeline provides great flexibility for researchers to initiate their research, and make use of the latest development in causal inference and machine learning.

We present three empirical studies conducted in distinct application contexts: an open-source software platform, an online microlending website, and a ride-sharing application. While they each start at a different stage along the pipeline, collectively, however, they demonstrate the effectiveness and flexibility of the proposed end-to-end pipeline in promoting pro-social behaviors.

CHAPTER I

Introduction

In his book *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Nobel Prize winner Richard Thaylor introduces the concept *nudge* as a behavioral mechanism that "alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives." The interventions are usually cheap and easy, but the achieved difference in people's behavior can be huge. In some of the most prominent applications, nudge practitioners changed the "default choice" and designed better-structured options, which successfully nudged participants to increase retirement savings, opt-in for organ donation, and make other pro-social choices.

Can we design better nudges, especially now that so-called "big-data" is deeply interwoven into our daily lives? In fact, many of our choices are already "data-driven," whether intentionally or not. Intentionally, we ask our phones about the weather, we search Google for answers, and we browse the product reviews. Unintentionally, however, many of our behaviors are actually shaped by data science algorithms, especially when we are interacting with online platforms. News websites, social media applications, and e-commerce platforms determine which headlines they want us to read, which news feed they want us to view, and which products they want us to shop, all in data-driven approaches, personalized through recommender systems or other filtering algorithms. And they have proven the effectiveness of this approach with their considerable revenue outcomes.

In observing the great potential of nudges in improving societal benefits, and the great success of data-driven business applications, this dissertation considers how might we unleash the power of data science to provide better nudges for pro-social behaviors. Or, more specifically, it asks can we use data science to better nudge pro-social behaviors?

Indeed, online platforms have provided us unprecedented opportunities for the nudge, with three highlights: First, online social media and web-based applications track and host fine-granularity data generated by real-world users. And the recently developed data mining algorithms and machine learning models provide tools to process and make sense of the data. Second, nudge theory's core concept is *choice architecture*, which refers to the context in which people make decisions. This particular term is in perfect analogy to the idea of recommender systems on the online platforms, which, in a general sense, are designed to facilitate users making choices. Lastly, interactive interfaces and rich functionalities make online platforms an ideal place to implement and deliver nudges to users.

1.1 Overview of an End-to-End Data Science Pipeline

With these three factors combined, we propose an end-to-end data science pipeline to identify, implement, and evaluate nudges that promote pro-social behaviors. As illustrated in Figure 1.1, such a pipeline consists of three critical stages:

Empirical data analysis We start by analyzing empirical data generated by people in real-world contexts. The goal of this stage is to reveal causal insights from user behavior data and identify potential "nudges" that may lead to data-driven solutions.

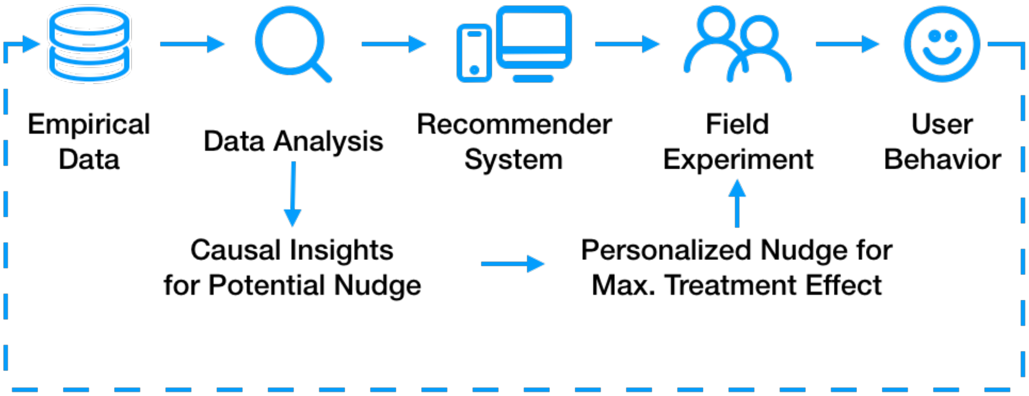

Empirical Data
Data Analysis
Recommender System
Field Experiment
User Behavior
Causal Insights for Potential Nudge
Personalized Nudge for Max. Treatment Effect

Figure 1.1: The End-to-End Data Science Pipeline

Recommender system As the second step, we design recommender systems that provide personalized suggestions for people to take action. The recommendations are personalized so as to maximize the nudging effect in changing people's behaviors.

Field experiment Finally, we implement and evaluate recommender systems. Note that it is best to deploy such systems in the original context, in order to demonstrate the effectiveness of the data-driven nudge in real-world scenarios.

These three stages comprise an end-to-end data science pipeline, with human-generated data on one end, and the change of human behaviors on the other. Note that the pipeline is not merely combining three separate types of data science applications. Instead, the three stages are closely knitted together and span the entire lifecycle of data, in which the output from upstream stages serves as input for downstream stages.

1.1.1 Challenges and Opportunities in each Stage

The connection between different stages goes even beyond the input-output coupling. In fact, we have to "re-purpose" the objectives of individual stages in order to build such an end-to-end pipeline. This gives rise to a series of challenges that call for interdisciplinary responses. As a result, this approach provides a principled framework that unifies efforts across different literature, even as it advances each component of the pipeline in new directions.

In this section, we re-examine each stage, highlighting challenges and opportunities separately:

Our first stage relies on a causal inference to draw causal conclusions from observed data, yet it is hard to apply such techniques directly when the data are massive, heterogeneous, and even unstructured. This challenge is to be addressed by developing new methodologies to conduct causal inference with machine learning algorithms. Indeed, researchers in computational social science are applying large-scale data analysis techniques to study problems with social, political, and policy implications. Many studies in this field derive causal conclusions, which work well as the first stage in our pipeline. In return, the pipeline allows downstream stages to turn the causal conclusions into actions, which can significantly increase the applicable value of the conclusions.

Machine learning has been widely used in developing recommender systems, which are optimized to predict user choices. To adopt a recommender system as the second stage of the pipeline, however, we need to change the optimization goal of the recommender system to optimize the treatment effect. That is, we want the recommender system to suggest personalized treatments that are most effective in nudging pro-social behaviors. Recently, researchers in the reinforcement learning and recommender system area are paying increasing attention to estimating heterogeneous treatment effects through "counterfactual learning" and learning optimal policy functions based on observed data. These studies closely resemble the second stage in our pipeline, especially if the objective of the policy is to maximize the "change" in users' adopted behavior. In return, the first stage of our pipeline can provide domain knowledge and theoretical guidance to improve policy learning. Also, the overall context of nudging pro-social behaviors can help the developed methodologies achieve a greater societal benefit.

The third stage in our pipeline is akin to the randomized field experiments used by behavior economists and other empirical researchers to evaluate new programs or policies. One common concern in field experiments is *non-compliance* among participants. As the third stage of the pipeline, however, the concern can be eased as the recommender system in the previous stage serve personalized treatments that are predicted to increase compliance and maximize the treatment effect. Traditionally, the analysis of randomized field experiments is focused on average treatment effect. Yet the same concern of non-compliance also highlights the importance of analyzing heterogeneous treatment effects.

The analysis of heterogeneous treatment effects calls for more sophisticated causal inference tools, which we have highlighted as an important contribution of the first stage. This implies that the pipeline is not a one-way street, but a loop.

1.1.2 Connecting End-to-End as a Loop

As we have seen, randomized field experiments are in need of advanced causal inference techniques, which are developed in the first stage of the pipeline. In the first stage, we want more empirical data to mine better nudges, but, if we treat the experiment data as a new set of human-generated data and feed it into the first stage, we can improve both stages at the same time. This essentially connects the two ends of the data science pipeline to form a loop! This allows us to repeat iterating the pipeline, each time with an updated collection of data, new insights for designing nudges, better-personalized recommendations, and more effective nudge implementation in the application.

Until now, illustrations of the data science pipeline always start with observed data collections and end with a new set of user behavior data. For this reason, the looping structure is a vast improvement, as it allows great flexibility in deciding where to start and end the loop. One may start with causal insights, such as insights from theory or previous literature, that are not necessarily discovered through a first stage, and proceed to develop recommender systems with machine learning algorisms. Instead of data-driven recommender systems, one may also start with a variety of rule-based recommender systems and conduct field experiments. If desired, one may decide to later update recommender systems using the experimental data. Similarly, researchers can choose to stop after one, two, or more stages depending on the intermediate results or external collaborations without worrying about the study's intactness. Such flexibility is evidenced by the three research projects in this dissertation. Each of them starts at a different stage in the pipeline.

Finally, we should note that the term "end-to-end" is different from the typical machine learning literature definations. In the machine learning context, end-to-end refers to models (especially deep neural network models) that directly transform raw inputs (such as text and audio/video streams) into target output (prediction labels, generated text) without explicitly specifying intermediate stages such as feature extraction, selection, and/or normalization. End-to-end machine learning models emphasize the ability to encapsulate a series of representations and transformations of intermediate data, once the input and output of a model are defined. However, the end-to-end in our proposed data science pipeline goes beyond individual machine learning models. Instead, it emphasizes how different data science methods, including causal inference, recommender systems, and field experiments, can be joined to cover the entire lifecycle of user-generated data and achieve pro-social benefits.

In addition, in our usage, end-to-end underlines the role of human participation in the process. End-to-end machine learning tries to minimize human effort in the training process by elimininting intermediate stages, so that only the raw input and target output are required. Our proposed end-to-end pipeline, however, is designed with humans at the center. Input data is collected from humans in real-world settings, and the output of each stage includes insights of human behaviors and personalized treatments. Finally, the success of the pipeline is also evaluated based on user behaviors observed in their original settings.

1.1.3 Real-World Applications of the End-to-End Pipeline

In this dissertation, we present three research projects that employ the pipeline in real-world applications with the goal of nudging pro-social behaviors. We can broadly refer to pro-social behaviors as any social behaviors that benefit other people or society as a whole. However, we bind the scope of this dissertation to pro-social behaviors on online platforms, where data-driven approaches are most feasible. Further, we focus on online platforms where the participating and contributing behavior understood to be pro-social. Note that we do not require such behaviors to be purely altruistic. Indeed, the three platforms studied in this dissertation range from a non-profit microlending website (Kiva, Chapter IV) to a commercial ride-sharing application (DiDi, Chapter V).

A summary of the three projects and an outline of the dissertation follows.

1.2 Dissertation Outline

In Chapter II, we review preliminaries of the causal inference literature. A recurring theme of our research is the interplay of causal inference and machine learning. As introduced in §1.1.1, causal inference plays important roles across different stages in our proposed pipeline, both in the experiment and non-experiment settings. Yet our work extends beyond the classic causal inference literature, as the data involved in the studies are massive, heterogeneous, and even unstructured, which are more easily handled with machine learning algorithms. In this chapter, we categorize causal inference literature with identification strategies and highlight recent explorations in bridging causal inference and machine learning.

Chapters III to V introduce three research projects that apply the end-to-end data science pipeline. On the surface, the three projects are independent of each other: they are conducted in distinct application contexts: an open-source software platform, an online microlending platform, and a ride-sharing platform; and they each start at a different stage along the pipeline. Collectively, however, they demonstrate the advantage and flexibility of the proposed end-to-end pipeline in promoting pro-social behaviors.

In Chapter III, we study the usage of emojis on GitHub. GitHub is the largest platform for open-source software development. A large proportion of conversations on GitHub are organized through its issue tracking system. Adequate and timely response to an issue is critical for the solution of the problem and the improvement of the project. We would like to show that using emojis for issues on GitHub projects attracts more attention from other developers, and leads to a faster resolving of the issue. This study is aligned with the first stage in our pipeline. It joins recent research on emoji usage on online platforms and is motivated by discussions of the potential benefit of emojis as non-verbal cues in facilitating online communications. By quantifying the effect of using emoji in online discussions, we identify a nudge that can potentially promote pro-social behaviors. We do not address downstream stages; however, a suitable response could include an emoji recommender system to encourage the use of emojis.

In Chapter IV, we develop a team recommender system for lenders in Kiva.org and evaluate it with a large-scale field experiment. Kiva is an online microlending platform that connects citizen lenders with low-income entrepreneurs in developing countries, and its continuing success relies on the active participation of its lenders. We build upon the empirical evidence that team identity promotes contribution to the public good, and leverage the power of machine learning to develop a recommender system that predicts new team membership based on historical data. Through a large-scale field experiment, we show that team recommendations can be an effective and low-cost behavioral mechanism to increase charitable contributions. This study spans the second and third stages of the pipeline. A companion study, which also studies team identity on Kiva.org, serves as the first stage for this study, as it reveales the causal insights of leveraging social identity to promote lending. [32] Evaluating recommender systems through a large-scale field experiment offers direct evidence of the power of the end-to-end pipeline in promoting pro-social behaviors in the real world.

In Chapter V, we examine the effect of team formation and inter-team contests on the productivity of DiDi drivers through a large-scale field experiment. DiDi is the dominant ride-sharing platform in China. Yet similar to other gig-economy platforms, its workers often find themselves lonely and disengaged, citing a lack of work identity and bonds with co-workers, which affects their productivity and their satisfaction with the job. In this study, we randomly assign drivers to teams based on different team formation strategies and have these teams compete for cash prizes. The experiment results verify the effectiveness of the team contest in increasing driver productivity, with a much larger effect for responsive teams. The results also suggest that the team formation matters, as teams comprised of drivers from the same region are more responsive and communicative within their team prior to the contest. Compared with the previous two chapters, we do not have the empirical data to support causal discoveries. Nor do we have behavior data to build a recommender system with machine learning algorithms. Instead, we begin with the third stage in the end-to-end pipeline. However, the data collected through the experiment enable us to learn the effectiveness of different team formation strategies and extend a team recommender system for further studies.

Finally, in Chapter VI, we conclude the dissertation and indicate directions for further exploration.

CHAPTER II

Preliminaries on Causal Inference

As we have seen in the introduction, causal inference plays several critical roles in our proposed end-to-end data science pipepline. In the first stage, we need causal inference to identify potential leverages that promote pro-social behavioral changes. In the third stage, we rely on randomized experiments to evaluate their promotion. Not only do we need properly designed randomized experiments, but we also them rigorously analyzed in order to demostrate their effectiveness in nudging pro-social behaviors, both of which demand guidance from causal inference literature.

However, our work extends beyond classic causal inference in that we are utilizing data of a much higher dimensionality. This explosion in dimensionality brings challenges to causal inference techniques, but we believe that machine learning can at least help alleviate such issues.

In this chapter, we review literature on causal inference. We mainly follow Rubin's potential outcome framework to set up causal inference problems, and categorize the causal inference techniques based on their identification strategy. In particular, we argue that our work is: (1) direct evidence (in that it reveals great potential) that machine learning and causal inference can go hand in hand in achieving social impact; (2) an exploration of how machine learning and causal inference can be combined; and (3) indicative of the challenges to both machine learning and causal inference.

As a final note, while there are other related literatures specific to individual chapters and projects of the dissertation, such as literatures on emoji usage on social media, social identity theory, and contest theory, we refer to them in their corresponding chapters.

Before descussing different identification strategies, we begin with a review of Rubin's Potential Outcome Framework [64], also referred to as the Rubin Causal Model. This establishes the notation for the remainder of the chapter.

2.1 Rubin Causal Model

The Rubin Causal Model has two key elements: *potential outcomes* and the *assignment mechanism*. We will start by defining the potential outcomes.

We use a binary random variable, $d_i \in \{0, 1\}$, to denote the treatment status of the individual i. The individual is treated iff $d_i = 1$ and untreated iff $d_i = 0$. We denote the outcome of the individual by y_i. For each individual i, we also observe a set of covariates, denoted as a covariate vector $\boldsymbol{x}_i$. The causal question of interest is whether y_i is *affected* by the treatment d_i.

To answer this question, we assume that we can imagine what the outcome might be if the individual is treated and if not. Hence, there are two potential outcomes for each individual: $y_i^{(1)}$ and $y_i^{(0)}$. $y_i^{(1)}$ is the outcome had the individual not been treated, regardless of his actual treatmment status, and $y_i^{(0)}$ is the outcome if the individual is treated.

In reality, we can only observe one of the two potential outcomes based on the treatment status. The observed outcome can be expressed as:

$$y_i = y_i^{(d_i)} = y_i^{(1)} d_i + y_i^{(0)}(1 - d_i) = \begin{cases} y_i^{(1)} & \text{if } d_i = 1 \\ y_i^{(0)} & \text{if } d_i = 0 \end{cases}. \tag{2.1}$$

The causal effect of the treatment on an individual, also referred to as individual treatment effect (ITE), can be represented by the difference of the potential outcomes:

$$\mathrm{ITE}_i = y_i^{(1)} - y_i^{(0)}. \tag{2.2}$$

The econometric literature has usually focused on the average causal effect of the treatment. If we average ITE over all individuals, we get the average treatment effect (ATE):

$$\mathrm{ATE} = \mathbb{E}[y_i^{(1)} - y_i^{(0)}]. \tag{2.3}$$

Alternatively, if we average the treated individuals, we get the average treatment effect on the treated (ATET, or ATT):

$$\mathrm{ATT} = \mathbb{E}[y_i^{(1)} - y_i^{(0)} \mid d_i = 1]. \tag{2.4}$$

Recently, with the help of machine learning algorithms, researchers have been giving increased attention to the heterogeneity in the treatment effect (with respect to observed covariates), and studying the heterogeneous treatment effect (HTE) in different subgroups of individuals [14].

Although we can observe the treatment an individual receives and the corresponding outcome for that individual, we cannot observe outcome of treatment that the individual does not receive, which is referred to as *counterfactual* outcome. Therefore, we cannot directly observe the causal effect, which is called the "fundamental problem of causal inference" [61]. Ultimately, causal effects can only be estimated by comparing different individuals under different treatments.

Indeed, one may easily calculate the observed difference between treated and untreated individuals as $\mathbb{E}[y_i|d_i = 1] - \mathbb{E}[y_i|d_i = 0]$, which can be rewritten as:

$$\begin{aligned} &\mathbb{E}[y_i^{(1)} \mid d_i = 1] - \mathbb{E}[y_i^{(0)} \mid d_i = 0] \\ =&(\mathbb{E}[y_i^{(1)} \mid d_i = 1] - \mathbb{E}[y_i^{(0)} \mid d_i = 1]) + (\mathbb{E}[y_i^{(0)} \mid d_i = 1] - \mathbb{E}[y_i^{(0)} \mid d_i = 0]). \end{aligned} \tag{2.5}$$

The first term is the ATT defined in (2.4). However, the ATT differs from the observed difference by $\mathbb{E}[y_i^{(0)} \mid d_i = 1] - \mathbb{E}[y_i^{(0)} \mid d_i = 0]$. This term captures the difference in average $y_i^{(0)}$ between the treated and untreated individuals, and is the so-called *selection bias*.

The existence of selection bias hinders our ability to draw causal inference from observational data. In economics, researchers have developed various strategies to eliminate selection bias and identify causality. These strategies are frequently referred to as *identification strategies*. Different identification strategies rely on different assumptions of how treatment is assigned to each individual, namely the *assignment mechanism*. This is usually characterized as a function of potential outcomes and covariates. In general, assignment mechanism is divided into three classes. The first class is to use randomized experiments; the second assumes unconfoundedness; and the third includes all remaining mechanisms. We review identification strategies for the three classes in subsequent sections.

2.2 Causal Inference with Randomized Experiment

Randomized experiments have been called the "gold standard" for establishing causality. In a randomized experiment, some individuals are randomly selected to receive treatment, while others remain untreated. The treatment assignment is independent of potential outcomes. That is, we get $\mathbb{E}[y_i^{(0)} \mid d_i = 1] = \mathbb{E}[y_i^{(0)} \mid d_i = 0]$. In this scenario, the selection bias term in (2.5) disappears, and the the observed mean difference between treatment and control groups is the unbiased estimate of the causal effect. Besides *completely randomized experiments*, there are also variations on randomization, such as *stratified randomization* and *pairwise randomization* [78]. But, regardless of the randomization, the probability of assignment can still be described as a function of the observed covariates and is independent of the outcomes.

Randomized experiments in evaluating programs have traditionally been rare [64], due to the high cost of implementing the interventions and tracking the subjects before and after the intervention. The rise of online communities, however, presents a practical and cost effective opportunity for conducting large-scale, randomized experiments [35]. The Internet and web-based applications offer an extended collection of technologies for intervention, such as sending emails or texts [32, 33], modifying web interfaces or application fuctionalities [20, 19], and using bots or other automated functionalities [41]. Sometimes it involves minimal collaboration with the site owners. For example, experimenters can register as regular users and design interventions without extra permissions or changes to the existing system [34, 32]. Alternatively, researchers can collaborate with the site owner or even deploy their own sites. This allows for more flexible intervention mechanisms and better access to data.

Although randomized experiments serve as the "gold standard" in testing causality, they do not guarantee that the correct causal effect is measured. We need to take precautions to ensure the randomized experiments is identifying the desired causal effect. One precaution is to design appropriate control conditions, which can be challenging. Take the analogy of a medical experiment. We need to have a *placebo* that resembles the stimulus except for the "hypothesized active ingredient." Such a placebo condition is also required for randomized experiments in other contexts.

The other precaution to be aware of is user's noncompliance. That is, the users' *actual* treatment status might be different from their *assigned* treatment status. In such cases, we often need to analyze the experimental data with *intend-to-treat* (ITT) analysis, and report the local average treatment effect (LATE). See 2.4.1 for a related discussion.

In sum, randomized experiments serve as a crucial step in our proposed pipeline. In this dissertation, we run randomized experiments to verify the effectiveness of the recommender system in changing users' behaviors, and we do so by collaborating with industrial partners and implementing new functionalities on their platform. The treatment we assign to each treated individual is personalized through the recommender system. We follow the literatures detailed above in taking all necessary precautions to ensure correct identification. However, our work extends this research by showing that such personalized treatment increases compliance.

2.3 Causal Inference under Unconfoundedness

In many cases, randomized experiments may not be an option, and we hope to derive causal conclusions based on observational data. Luckily, there is mature literature on estimating average treatment treatment effect if the *unfoundedness* assumption can be justified.

Unfoundedness assumes that we can observe all the confounders, factors that are associated with both the treatment assignment and the potential outcomes. Therefore, we can assume that the treatment assignment is independent of the potential outcomes conditional on observed confounders. This assumption is also referred to as the *conditional independency assumption* (CIA), and can be written as:

$$(y_i^{(1)}, y_i^{(0)}) \perp\!\!\!\perp d_i \mid \boldsymbol{x}_i. \tag{2.6}$$

Under unfoundedness, the treatment assignment is "as good as" random, and the observed difference between the treated and untreated individuals who share the same values as the confounders can be interpreted as the causal effect [101]. In fact, randomized experiments can be regarded as a special case of unconfoundedness where the treatment assignment is indeed random.

Note that the underlying presumption of "observing" any difference is that there has to be at least one treated and one untreated individual. This requirement can be written as:

$$0 < \mathrm{P}(d_i = 1 | \boldsymbol{x}_i = \boldsymbol{x}) < 1 \quad \forall \boldsymbol{x}, \tag{2.7}$$

and is called the "overlap" or "common support" assumption. It implies that the support of the conditional distribution of $\boldsymbol{x}_i$ given $d_i = 1$ overlaps with that of the conditional distribution of $\boldsymbol{x}_i$ given $d_i = 0$.

When the overlap assumption is satisfied, we can estimate the *conditional average treatment effect* (CATE) as:

$$\begin{aligned}\mathrm{CATE}(x) &= \mathbb{E}[y_i^{(1)} - y_i^{(0)} \mid \boldsymbol{x}_i = \boldsymbol{x}] \\ &= \mathbb{E}[y_i^{(1)} \mid \boldsymbol{x}_i = \boldsymbol{x}] - \mathbb{E}[y_i^{(0)} \mid \boldsymbol{x}_i = x] \\ &= \mathbb{E}[y_i^{(1)} \mid \boldsymbol{x}_i = \boldsymbol{x}, d_i = 1] - \mathbb{E}[y_i^{(0)} \mid \boldsymbol{x}_i = \boldsymbol{x}, d_i = 0] \\ &= \mathbb{E}[y_i \mid \boldsymbol{x}_i = \boldsymbol{x}, d_i = 1] - \mathbb{E}[y_i \mid \boldsymbol{x}_i = \boldsymbol{x}, d_i = 0].\end{aligned} \tag{2.8}$$

There are a variety of methods that aggregate CATE to derive the average treatment effect, most of which use *regression*, *matching*, *propensity score*, or combinations of these three methods.

Regression If we assume a linear relationship between the outcome and the covariate, that is, $y_i^{(0)} = \beta_0 + \beta \boldsymbol{x}_i + \epsilon_i$, the observed outcome can be writen as follows:

$$\begin{aligned}y_i = d_i \cdot y_i^{(1)} + (1 - d_i) \cdot y_i^{(0)} &= d_i \cdot [y^{(1)} - y^{(0)}] + y^{(0)} \\ &= d_i \cdot \mathrm{ITE}_i + \beta_0 + \beta \boldsymbol{x}_i + \epsilon_i\end{aligned} \tag{2.9}$$

If we assume ITE_i to be constant, that is $\text{ITE}_i = \text{ATE}$ for all i, we can fit a regression model to estimate the average treatment effect. Despite the simplicity, however, regressions come with a few caveats. First, regression assumes linearity, which may lead to inconsistency if not hold. Although one can incorporate non-linear terms into the regression formular, in general, it is hard to determine the right function forms. Second, regression does not account for the covariate distributions of the treated and untreated, and it will not indicate if there is no common support. This may lead to a poor estimation of counterfactual outcomes [106]. Finally, regression methods estimate ATE rather than ATT, but in many scenarios, ATT is a much more interesting estimator than ATE.

Matching With (2.8), the treatment effect can easily be calculated by matching treated and untreated individuals with the same covariate. In practice, however, the overlapping assumption may not hold at every $\boldsymbol{x}$, especially when the dimensionality gets higher. This is known as the *curse of dimensionality.* As a result, many individuals may not be matched with individuals of opposite treatment status. Therefore, inexact matching is usually required instead of exact matching. With inexact matching, we are estimating the term $\mathbb{E}[y_i \mid \boldsymbol{x}_i = \boldsymbol{x}, d_i = 0]$ in (2.8) as:

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i = \boldsymbol{x}, d_i = 0] = \sum_{j, d_j = 0} w(i, j) y_i, \tag{2.10}$$

where $w(i, j)$ is the weight of how the individual j contributes in constructing counterfactual outcomes of i.

The most common way to do inexact matching is to define a similarity or distance function and then match each treated individual to the nearest untreated individual(s), also known as *nearest neighbor matching.* In this case, $w(i, j) = 1$ if and only if j is the closest untreated individual to i. Similarly, k-nearest neighbour matching would mean $w(i, j) = 1/k$ if j is among the k closest untreated individuals. Other inexact matching methods, such as kernal matching and local linear matching, can also be specified with variations of $w(i, j)$.

None of the matching methods can circumvent the common support assumption, as it is always likely that some treated individuals are far from any untreated individuals. Unlike regression, however, matching can at least highlight common support problems, and researchers can decide how to handled unmatched or ill-matched cases.

Instead of finding matches in the original high-dimensional space, an alternative method estimates a *propensity score* for each representation and performs matchings on the scores. Below, we first introduce propensity score and its properties, and then discuss propensity score matching and other applications of propensity scores.

Propensity Score Propensity score is defined as the conditional probability of being treated given the observed covariates. In notation, the propensity score, $e(x)$, can be written as:

$$e(x) = P(d_i = 1 \mid \boldsymbol{x}_i = \boldsymbol{x}) = \mathbb{E}[d_i \mid \boldsymbol{x}_i = \boldsymbol{x}]. \tag{2.11}$$

The simple way to estimate the propensity score is to compute the proportion of treated individuals for each cell defined by the covariates. However, this can create the same curse of dimensionality. In practice, the propensity score is usually estimated using a parametric model, such as probit or logit regression.

[101] shows a nice property of the propensity score. That is, under unconfoundedness, the independence of potential outcomes and treatment status still holds after conditioning only on the propensity score. Mathematically,

$$d_i \perp\!\!\!\perp (y_i^{(1)}, y_i^{(0)}) \mid \boldsymbol{x}_i \Rightarrow d_i \perp\!\!\!\perp (y_i^{(1)}, y_i^{(0)}) \mid e(\boldsymbol{x}_i). \tag{2.12}$$

Intuitively, this means that if the propensity score is correctly estimated, it should encapsulate all the information we needed from the covariates. For individuals of the same (or similar) propensity score, the difference in observed outcomes comes solely from differences in treatment. Therefore, we can match treated individuals with untreated individuals who have similar propensity scores. This method is known as *propensity score matching* (PSM). Similar to the more general matching introduced above, there are many variations of propensity score matching. In addition to nearest neighbour matching, we can also stratify the propensity score and match individuals sharing the same stratum (also known as interval matching).

In addition to matching, there is also mature literature on using propensity scores for weighting. We may rewrite 2.3 as:

$$\text{ATE} = \mathbb{E}[y_i^{(1)}] - \mathbb{E}[y_i^{(0)}] = \mathbb{E}\left[\frac{d_i \cdot y_i}{e(\boldsymbol{x}_i)}\right] - \mathbb{E}\left[\frac{(1-d_i) \cdot y_i}{1-e(\boldsymbol{x}_i)}\right]. \tag{2.13}$$

That is, we can re-weight each individual by the inverse of such probability and the difference between the weighted observations can be used to estimate the average treatment effect. This method is referred to as *inverse probability weighting* (IPW).

In general, *regression*, *matching*, and *propensity score* based methods are the most common tools in causal inference under unconfoundedness. It is also common to combine different approaches in practice, for example, doubly robust estimators that combines regression and propensity score methods have been proposed to increase the robustness to misspecification of parametric models [100].

As indicated earlier, randomized experiments also satisfy the unconfoundedness assumption. Therefore, the introduced methods – regression, matching, propensity score – can also be used to analyze data collected from a randomized experiment. This enables us to feed the output of the third stage into the first stage and transform our end-to-end pipeline into a loop.

Machine Learning "under Unconfoundedness" In recent years, more and more attention has been paid to using machine learning to help causal inference [15]; a majority of these assume unfoundedness. One stream of such literature focuses on the average treatment effect, where machine learning is mainly used to provide more flexible control for a large number of covariates [22, 38]. Another stream focuses on the heterogeneous treatment effect instead. Examples include [14, 115].

Alternatively, the causal inference techniques under confoundedness have also been applied to improve machine learning systems, such as search engines, recommender systems, and computational advertising [116, 82, 27]. Many such systems rely on implicit feedback from users as signals for training and evaluation, while many biases in human feedback need to be handled by causal inference models. A commonly studied question is off-policy evaluation, that is, "how will the performance improve if we change our system in this way." Without A/B testing, this question requires estimations which are "counterfactual" to what we can observe[117, 109]. This stream of literature is often referred to as counterfactual machine learning. We point to [68, 67] for more in-depth discussion.

Finally, a few impressions based on a preliminary review of these literature: The most popular means to adapt machine learning for causal inference are Lasso and Random Forest, while the most commonly borrowed causal inference techniques are propensity score based methods, especially weighting. This is not surprising: As a linear model, Lasso is not only simple, but also functions as a feature selection tools to eliminate excessive number of features. However, Random Forest is better at capturing non-linear interactions, and natually constructs partitions for matching. On the other hand, (propensity) scoring and weighting are easier to incorporate into existing machine learning models, and the estimation of scoring or weighting itself can easily be cast as another prediction problem.

2.4 Other Identification Strategies

So far, the causal inference techniques discussed in this chapter are based on the unconfoundedness assumption. In many scenarios, however, we cannot assume unconfoundedness. That is, there is still dependence between the treatment assignment and the potential outcomes conditioned on all observed confounders.

In such cases, none of the above methods can fully address the selection bias, and there is no general solution. However, researchers have identified a few special cases where solutions are available with additional assumptions. The most commonly known approaches are instrumental variables, regression discontinuity, and difference-in-differences.

2.4.1 Instrumental Variables

An *instrumental variable*, denoted as z_i is a variable that satisfies two criteria:

Partial Correlation The instrumental variable is partially correlated with the causal variable of interest, which is d_i in this case.

Exclusion Restriction The variable is not correlated with other determinants of the outcomes. The intuition is that, z_i is correlated to the outcome variable only through its partial correlation with the treatment.

If both criteria are satisfied, one can use *Two Stage Least Square* (2SLS) to derive an unbiased estimate of the treatment effect. Note that although the partial correlation criterion can be calculated and evaluated, the exclusion restriction criterion must be argued on a case by case basis. Some of the most widely known examples of instrumental variable studies include [12, 11], both of which use date of birth, an exogeneous variable beyond the control of each individual, as the instrument variable for the treatment assignment.

Instrumental variables can also be applied in combination with randomized experiments. Recall from §2.2 that when non-compliance exists in randomized experiments, the realized treatment status is confounded, as not all participants comply with the treatment status based on some endogeneous factors. In such cases, the *assigned* treatment can be used as the instrument for the *realized* treatment. It is obvious that the two are partially correlated. In addition, exclusion restriction can be justified as the treatment assignment is random, and it affects the realized outcomes of the subjects only through its correlation with the treatment status. This method is called the *intent-to-treat* (ITT) analysis.

We should be careful about the interpretation of the results from IV. If we assume that the treatment effect is constant, the estimate would be both ATE and ATT. If the treatment effect is heterogeneous, however, we can only obtained the *local average treatment effect* (LATE) [65]. As a special case of heterogeneouos treatment effect, LATE reports the average treatment effect on a subgroup of the subjects called *compliers.*

2.4.2 Regression Discontinuity

Regression discontinuity assumes that the treatment is determined by an observed forcing variable being on either side of a common threshold. The threshold creates a discontinuity in the conditional probability of treatment assignment. The forcing variable may be associated with the potential outcome, but such an association needs to be smooth. If a forcing variable and a discontinuity can be justified, we can assume individuals *close to* the boundary are similar not only in their covariates but also in their potential outcomes. For them, the treatment assignment can be seen as random and the difference in the observed outcomes would represent the treatment effect.

That is:

$$\begin{aligned}\mathrm{ATE}_{\mathrm{RD}} &= \mathbb{E}[y_i^{(1)} - y_i^{(0)} \mid v \approx c] \\ &= \lim_{v \to c+} \mathbb{E}[y_i \mid v_i = v] - \lim_{v \to c-} \mathbb{E}[y_i \mid v_i = v],\end{aligned} \tag{2.14}$$

where v_i is the forcing variable and c is the threshold.

Discontinuities are usually found when there is administrative policy that has transparent rules in assigning treatment. For example, in studying the electoral advantage to incumbency, [77] uses the vote share as the forcing variable and the 50% majority vote as the threshold.

2.4.3 Difference-in-Differences

The difference-in-differences (DID) method is frequently used to analyze natural experiments, where individuals are divided into treatment groups and control groups naturally by policy changes or natural phenomena, and we can observe the outcomes for both the treated and untreated individuals both before and after the natural experiment. Under such conditions, only individuals in the treatment group are exposed to treatment. In addition, they are exposed only after the treatment starts. We can first compute the difference within each individual before and after the time of the treatment, and then compare the differences among individuals across treatment groups. After the double differencing, the biases due to the permanent difference between the control and treatment groups and the biases due to the time change regardless of the treatment are both removed.

It is worth mentioning that the three strategies introduced in this section can be used either alone or in combination. For example, in analyzing randomized field experiments, one may use instrumental variables for intent-to-treat analysis, and at the same time construct counterfactual outcomes using difference-in-differences.

CHAPTER III

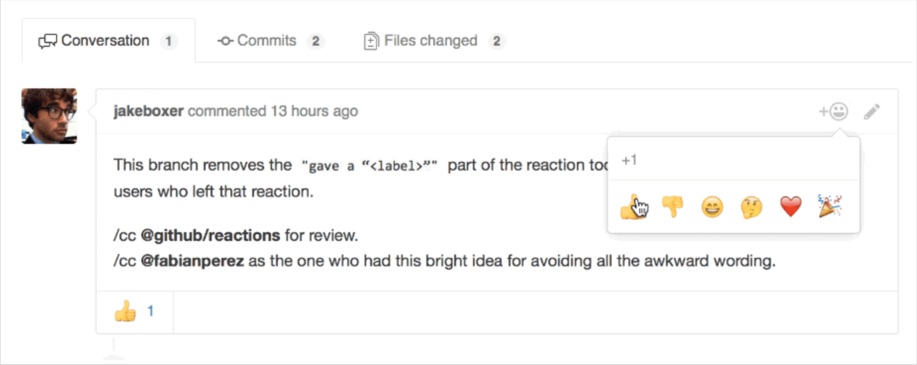
Emoji Promotes Developer Participation on GitHub

As demonstrated in Chapter I, the first step to promoting pro-social behavior is to understand what may affect people's behaviors. In this chapter, we present a study that focuses on this first stage of the pipeline. More specifically, we focus our study on developer participation on GitHub, the world's largest open-source platform. Through a careful statistical analysis, we show that the use of emojis in presenting an issue increases discussion participation. These findings not only deepen our understanding of developer communities, they also provide design implications on how to facilitate interactions and broaden developer participation.

3.1 Introduction

As the Linus's law of software development states, "given enough eyeballs, all bugs are shallow" [97]. That said there are never enough eyeballs in the developer community. On one hand, there might not be enough experts in the field. On the other, the such computer-mediated communications (CMC) may be less engaging in nature, due to a lack of *non-verbal cues*, which occur natually in face-to-face conversations [75]. These cues include body language, facial expressions, eye contact, vocal intonation, personal distance, etc.

An intuitive way to nudge more participation in the online developer community, therefore, is to bring non-verbal cues into the conversations. Indeed, Github implemented a "reaction" function in March 2016, in an effort to establish such non-verbal cues and facilitate communication between developers. Similar to the reaction function in Facebook, the "reaction" function allows users to select from a predefined set of emoji as a reaction to a conversation on GitHub, (as shown in Figure 3.1).[1] As of June 2019, GitHub supports eight reaction types, namely 👍 (+1), 👍 (-1), 😄 (laugh), 😕 (confused), ❤️ (heart), 🎉 (hooray), 🚀 (rocket), and 👀 (eyes).[2]



jakeboxer commented 13 hours ago

This branch removes the "gave a "<label>"" part of the reaction toc users who left that reaction.

/cc @github/reactions for review.
/cc @fabianperez as the one who had this bright idea for avoiding all the awkward wording.

Figure 3.1: A Screenshot of the Reaction Function on GitHub.

Beyond the eight reactions, however, emoji have been supported on Github since as early as 2014.[3] These graphic symbols carrying specific meanings are quickly adopted into online conversations, supported by multiple platforms, and inducted into Unicode standards. Indeed, in recent years, several researchers have focused on understanding emoji usage on online platforms, citing emoji as the the ideal non-verbal cues to express sentiment, strengthen expression, and adjust tone in online communication, where facial expressions or body gestures are not available [62].

[1] https://github.com/blog/2119-add-reactions-to-pull-requests-issues-and-comments

[2] The annotations for these emoji (in brackets) are provided by the GitHub Developer document (https://developer.github.com/v3/reactions/, retrieved in June 2019.)

[3] https://guides.github.com/features/mastering-markdown/, last updated on Jan 15, 2014, according to the web page, retrieved in June 2019

However, existing literature has not yet quantified the impact of using emoji on online platforms. In this study, we take the initiative to quantify the effect of emojis in promoting participation within the developer community. We ask: are emoji attracting more eyeballs and soliciting more contributions? This study contributes not only to the study of developer participation on Github but also to the literature on the provision of public good as well as the literature on emojis. Specifically, we hypothesize that:

H1: *Using emoji in an issue increases the participation of GitHub users in the conversation.*

The participation of users in a conversation may be measured in whether the issue gets commented on, the number of users commenting on the issue, or the number of comments per user.

Not only do we care about whether using emojis attracts more users to the discussion, we would also like to find out if the discussion is more likely to lead to a resolution of the issue, and if yes, whether the resolution is completed in a timely manner:

H2.1: *Issues with emoji are more likely to be resolved.*

H2.2: *Issues with emoji are resolved in a shorter time period.*

Besides the effect of emoji on the participation in and the outcome of the development task, we are also interested in whether the use of emoji in a conversation affects the culture of the developer community. In particular, whether it reshapes the norm of conversations on GitHub towards using more emoji:

H3. *Using emoji in an issue increases the use of emoji in the comments.*

The most straightforward way to test these hypotheses is to compare issues with and without emoji. Yet the vanilla t-test would be biased due to self-selection. Indeed, taking H1 as an example, there are many confounding factors that affect both the use of the emoji in an issue (the treatment) and whether that issue gets discussed (the outcome). For example, issues with emoji are usually posted on projects with higher fork numbers ($p < 0.001$), and these issues are more likely to receive comments just because the projects are popular. In this study, we implement a rigorous statistical method, namely propensity score matching (PSM) [101], to estimate the causal impact of using emojis in an issue. By using PSM, we isolate the confounding variables and find issues that differ only in whether they used emoji or not.

In the rest of this chapter, we review the related literature in 3.2, introduce the detailed setting and data set of this study in 3.3, illustrate our methodology in 3.4, and lay out the hypotheses to be tested in 3.5.

3.2 Literature Review

Our research is closely related to three streams of existing literature: emoji usage analysis, language style in online communities, and participation in open-source communities.

Emoji Usage Analysis Increasingly popular, emoji have almost become a ubiquitous language in recent years. Research on emoji usage has been conducted in a number of applications and scenarios, such as input methods [84], instant messaging apps [121], and social networks [21]. Compared to plain text, their compact visual representation has attracted researchers to study the intentions of using emoji [62] and the semantics and sentiments of emoji [52, 2]. Their rich semantics has also made emoji prone to misinterpretations and ambiguity, which is discussed extensively in [92, 91, 2]. These studies motivate us to study the community-specific properties and interpretation of emoji. To the best of our knowledge, this paper is the first to study emoji usage in a tech community and the first one that measures its effect in attracting participation in that community.

Language Styles in Online Communities To emoji or not is arelated to word choice in different language styles, which connects our work to literature on language styles in online communities. Language choice may affect the attractiveness of a message [111]. Thus, language style, together with how it evolves, is believed to be part of community norms [46, 45, 55]. Research both in lab settings [93] and on Twitter [44] has shown that participants tend to converge to its community's language style. We will show that emoji are part of the language norm on GitHub, and we will discuss how such a norm is formed.

Participation in Open-Source Communities Open-source communities and platforms, such as GitHub, have attracted the attention of researchers in various fields, such as software engineering [94, 105], CSCW [43, 113, 87], and management science [99, 17]. Researchers have identified many key factors that affect participation and performance, such as network structure among the collaborators [87] and status motivations [99]. Most of this work focuses on collaboration and coding performance. Our work focuses on communications through issues and pull requests, and analyzes the effect of non-verbal language tokens (emoji) on user participation.

3.3 GitHub and its Archival Data

Before offering our analysis, we first introduce background information about GitHub and how our data are collected and processed to enable analysis.

GitHub[4] is the largest host of source code in the world. It offers distributed version control and source code management via the Git protocal, and has become one of the most popular platforms for hosting open-source software.

On open-source software platforms like GitHub, most distributed development activities are coordinated through issues and pull requests. Issues can be posted by

[4] https://github.com/

any user to report software bugs, enhancement suggestions, or to solicit help. They are frequently used as a tracking system for ideas, enhancements, tasks, bugs, or other user feedback.[5] On the other hand, pull requests (Abbreviated as PR) are proposed changes to the code repository. Collaborators can discuss and review such changes and decide if the change should be merged into the code base.[6]

Conversations on GitHub are organized through comments on the issues and pull requests. Different from common online chatters, these conversations can directly influence the quality of the projects. Adequate and timely response to an issue is critical for the solution of the problem and the improvement of the project. Therefore, we focuses on issues and their responses in this project.

Although GitHub hosts both private and public code repositories, we only focus on the public repositories. These repositories are accessible to all users, and are considered "open-source."Activities on these public repositories, such as creating, closing, or commenting on an issue, are collected by GibHub, and streamed to the public through its *Events* APIs.[7] A third-party website, named GHTorrent,[8] monitors the public events streams and maintains a scalable, queryable, offline data mirror for public access. It also actively queries GitHub's API to retrieve profile information of both users and projects with an internal algorithm.

Our analysis is based solely on data hosted on GHTorrent. Specifically, we collect all the issues created in open repositories between January 1st, 2016 and June 30th, 2017, and we track their associated comments and closing events. In order to gather more contextual information, we also retrieve the user and repository profiles related to the collected issues. However, we acknowledge that GitHub does not provide backtracking for historical profiles and that GHTorrent retrieves and archives profile information at a self-determined frequency. Therefore, we use the archived profile

[5] https://help.github.com/en/articles/about-issues

[6] https://help.github.com/en/articles/about-pull-requests

[7] https://developer.github.com/v3/, retrieved June 2019

[8] http://ghtorrent.org/, retrieved June 2019. Also see [57].

Table 3.1: GitHub Data Collection on GHTorrent

Table	Description
Events	Public event streams, where `IssuesEvent` and `IssueCommentEvent` will be extracted.
Repos	Repository where an issue is posted.
Users	Users who create the issues.

information that is closest to the creation of an issue as the surrogate. The list of data collected in this study is summarized in Table 3.1.

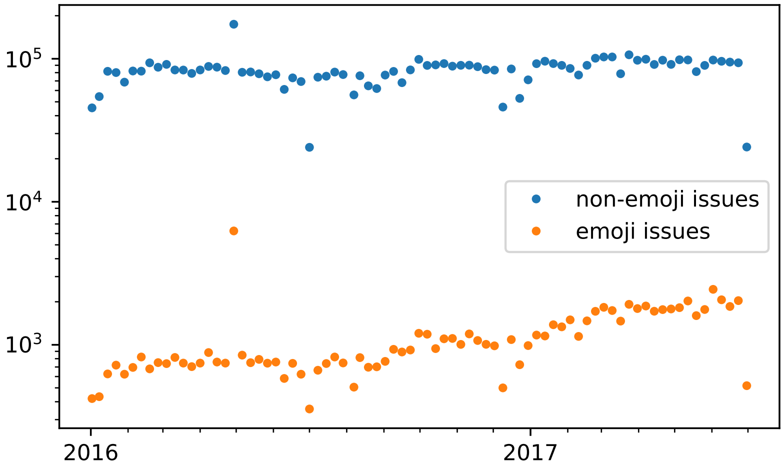


Figure 3.2: Weekly Stats of Issues with and without Emojis.

We plot weekly trending stats in Figure 3.2. Each dot represents the number of issues with or without emoji created in a week. The issues with emoji are still relatively few compared with issues without emoji. In fact, among the 11 million issues that we track, only 1.33% of them used one or more emoji. However, we do see an increasing trend in the use of emoji.

The imbalanced ratio between issues with and without emoji posts a class imbalance challenge for most machine learning models. Intuitively, a model would achieve $> 98\%$ accuracy by predicting all issues as not using emoji. To address the imbalance problem, we perform undersampling to match the number of issues with and without emojis at the week level before further analysis. After the undersampling, we arrived at a dataset with 366,382 issues posted by 204,265 authors in 165,969 repositories.

In the rest of the chapter, we examine if using emoji has a positive effect on the open-source platform.

3.4 Propensity Score Matching

We followed the Propensity Score Matching introduced in §2.3 to formulate the problem. A majority of the issues do not use emoji at all. The issues that did use emoji can be regarded as the early adopters from the perspective of innovation diffusion. Therefore, it is reasonable for us to focus on them first. This leads us to focus on the average treatment effect on treated (ATET, or ATT).

In our case, the propensity score is the probability that an issue uses emoji. By propensity score matching, we identify issues with a similar propensity score and assume that these issues are comparable. In such a way, we run a pseudo-randomized experiment in which the treatment of using emoji is randomly assigned to issues that are similar otherwise. The different outcomes of these issues are, therefore, only caused by whether they used emoji or not.

In classic econometrics literature, the propensity score is usually estimated with a logistic regression model, with treatment variable d as the dependent variable and the covariate X as the independent variables. However, the estimation is no different than a machine learning predictive model that learns to predict the treatment d with covariate X. In our analysis, we are going to apply two standard and commonly used machine learning algorithms to estimate the propensity score.

Below, we first detail the implementation of the propensity score estimation in 3.4.1 and assess the covariate balance in 3.4.2.

Table 3.2: Features used in propensity score estimation.

Category	Features
issue	length, posting time, text content (through topic modeling)
repository	# stars, # forks, # watch, # open issues, main program language, repository age,
issue author	# follower, # following, # public repos account age prior emoji usage.

3.4.1 Propensity Score Esimation

Although logistic regression is commonly used to estimate the propensity score, it assumes the linearity and additivity, which may or may not hold in reality. Instead of explicitly specifying non-linear terms (such as higher-order terms) or interaction terms in logistic regression, we apply a machine learning algorithm, namely Gradient Boosted Regression Tree (GBRT), as our propensity score model. This is aligned with recent literature [118, 76], which demonstrate with simulation studies that machine learning algorithms can improve the balance between treated and untreated groups.

The data collected from GHTorrent have provided us abundant information not only about the issues themselves but also about the context of the issues, such as the repository in which an issue is posted and the author who raised the issue. Such contextual information would also affect the probability of using emoji and should be modeled into the propensity score estimation. We summarize the features for propensity score estimation in Table 3.2.

As presented in Table 3.2, most features are structured, either as boolean/numerical variables or as categorical variables, the latter of which can easily be represented as a series of dummy variables. However, the text content of an issue is unstructured by nature. By all means, the topics expressed in an issue may correlate with emoji usage and must be accounted for in the propensity score estimation. For example, users who post issues to solicit help may be more likely to use emoji to express their sentiment, or users who report bugs may be more likely to use certain emoji (such as

🐛) to refer to bugs, etc.

The simplest way to represent text as structured features is to use the bag-of-word representation, which treats each word (or other language units) as a feature and its occurrence (sometimes reweighted with TF-IDF) as the value. Although this approach has been successful in many information retrieval models and text mining applications, it usually explodes the dimensions of the feature representations and increases the sparsity in the data, leading to slow convergence or even non-convergence.

There are several ways to reduce the dimensions of the text features. For example, one may use word embedding techniques (such as word2vec [90] or LINE [112]) to tranfer each word into a vector in a low dimension space. Each issue can be represented as a vector in the same space by aggregating its words in a pre-defined way. However, it is usally hard to interpret the meaning of each dimension. In this work, we apply a principled machine learning algorithms to cluster the text into a smaller number of topics, which are commonly referred to as topic modeling [25]. Specifically, we apply the Latent Dirichlet Allocation (LDA) model, which has been widely adopted in text mining applications.

Topics in Issues As previously outlined, we conduct topic modeling using LDA before training propensity estimation models in order to obtain the topic representation of the issue text. The number of topics is usually determined by trying different numbers, interpreting the word distributions of the topics, and selecting the models that are most interpretable.We empirically set the number at 30, which performs well in separating different topics. We present the discovered topics in Table 3.3, where we show the representative words in each model and the topic labels heuristically assigned based on the representative words. We can observe that several topics have a clear meaning.

Table 3.3: Topics in the Issues

ID	Topic	Representative Words
0	Ruby (package)	fastlane, version, users, ruby, gems, ios, end, build, xcode, library, false, app, http-cookie, lib, env
1	Ruby/Game	server, client, thread, log, understood, item, message, address, export_method, received, here's, sending, handler, mod, sent
2	GitHub (bug/request)	x, please, version, issue, bug, report, check, feature, information, z, expected, issues, api, pod, one
3	PHP	var, function, php, object, task, http, array, exception, diff, web, line, virtualenv, app, vendor, null
4	code/mixed	value, type, data, pass, released, left, invisible, name, distant, vendors, integer, green, red, record, rouge
5	code	e, de, r, b, c, u, n, l, w, p, v, la, en, x, que
6	Python	file, debug, line, python, lib, usr, couple, error, root, self, local, home, py, docker, fail
7	Rust (language)	src, group, future, build, png, integrate, cargo, mins, go, core, frustrations, rustc, home, downloading, panicking
8	code (keyword)	use, code, string, using, test, example, new, return, type, public, like, project, data, one, get
9	C/Java	src, h, error, include, home, c, int, const, usr, jdk, function, warning, future, void, local
10	JavaScript	error, node_modules, users, js, app, build, version, node, code, npm, get, module, lib, run, import
11	help request	would, like, use, new, using, png, one, work, add, get, could, also, user, see, make
12	Roda (Ruby library)	f, instructions, commercial, error, warning, game, sound, c, roda, earlier, engine, win, games, documents, users
13	JavaScript	silly, c, active, users, packages, error, verbose, atom, core, facility, node_modules, program, npm, files, lib
14	Gulp (JavaScript tool)	compiling, gulp, android, decoded, go, src, build, storage, debug, turn, users, package, h, detecting, ctx
15	code/mix	system, library, debug, managing, future, thread, lo, frame, lib, usr, interact, selects, panicking, context, layout
16	Node.js (JavaScript library)	npm, err, depth, common.py, support, number, version, add, theme, index, please, data, leaks, product, according
17	Java	info, c, error, test, source, main, java, users, failed, ago, file, jar, researcher, class, method
18	Swift/IDE	g, nil, let, workaround, spacemacs, variables, emacs, file, vim, branch, window, behaviour, layers, setup.py, join
19	HTML	align, supports, td, center, aliases, right, option, value, class, implication, width, height, l, codecs, kanji
20	code	name, id, map, xml, event, key, nil, values, layer, select, progress, plugin, data, public, resource
21	code (html)	class, div, style, width, href, color, title, faraday-cookie_jar, text, img, src, image, alt, css, height
22	MacOS	usr, local, bin, install, git, build, version, library, homebrew, checking, installing, package, remote, directory, installed
23	code	false, n, true, user, text, name, given, time, z, f, place, say, ti, target, reach
24	GitHub	issue, add, github, close, update, create, comment, list, column, issues, project, columns, pull, default, card
25	code/mix	tests, test, first, missing, ok, found, flow, vom, cpu, workstation, device, line, memory, rb, devices
26	Visual Studio/C++	studio, ptr, vs, target, href, zombies, app, plants, player, worker, video, id, accordance, multiply, mail
27	Ruby	lib, gems, ruby, bundle, unit, block, usr, vendor, users, home, uploaded, call, local, bundler, opt
28	contact/logistics	br, windows, href, pdf, android, download, free, experimental, performing, ordered, fault, partially, word, pinned, mailto
29	Rake (Ruby tool)	rake, version, browser, behavior, description, steps, reproduce, windows, url, system, mobile, problem, operating, expected, type

Evaluation Generally, machine learning tasks are evaluated with out-sample prediction accuracy. People usually adopt cross-validation to train their models on the training set and evaluate their models on a hold-out validation set. This simulates the application scenario where the models trained on the labeled training set are used to predict the labels on the unseen test set.

In propensity score estimation, however, out-sample prediction accuracy is not the focus of the prediction, as we already know the treatment of each sample. Indeed, there is not a single metric that can be used to evaluate the propensity score. The reason is simple: In observational studies, we would never know the real probability of getting treated. The lack of ground truth prevents any objective measurements of the accuracy of the predicted propensity score. Instead of an accuracy score, however, we can evaluate whether the matching based on the estimated propensity score is good enough to derive causal conclusions. Since there are several variations of propensity score matching, the evaluation method depends largely on the matching. In the next subsection, we first describe the matching method and then perform alternative evaluations on propensity score estimation.

3.4.2 Propensity Score Stratification and Balance Check

Now that we have estimated the propensity score for each issue, there are several ways to match issues on their propensity scores, such as nearest neighbor matching, kernel matching, and stratification matching. Before making decisions on the method to use, it is helpful to perform a visual analysis on the distribution of the propensity score. In Figure 3.3, we stratified the issues into 20 equal-size strata based on their propensity scores. For each stratum, we plot the average propensity score of all issues and the the true selection ratio (ratio of issues with emoji). If the propensity score is estimated correctly, the selection ratio of each stratum should equal the mean propensity score of the apps in the bin. In other words, we should expect a line close to $y = x$. Indeed, the blue dotted line in Figure 3.3 is aligned with the reference diagonal line (orange dashed), which is reassuring.

However, a concerning fact shown in the figure is that for the last two strata, the ratios of issues with emoji are very close to 1 (0.983 and 0.999), which poses a large threat to the *common support* assumption introduced in §2.3. That is, there are too few untreated samples to calculate the difference between the treated and untreated samples. [107] suggests a trimming procedure to exclude the intervals which lack common support. As we will see later, these two strata fail the balance check, which is also likely due to the violation of the common support assumption.

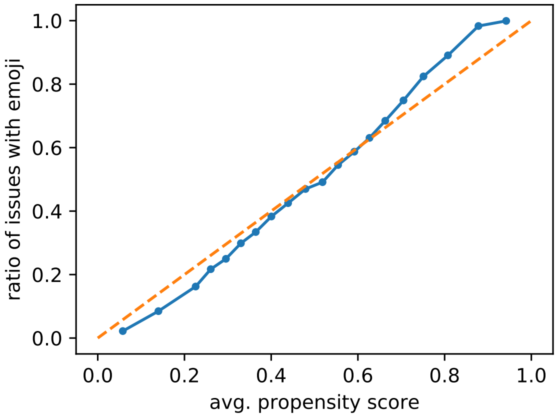


Figure 3.3: Binned average propensity score and true selection ratio

The trimming procedure natually leads us to the stratification matching, which is also referred to as interval matching, blocking, and subclassification [102]. In our case, we reuse the 20 equally sized strata based on their propensity scores. In such a way, each stratum has a similar number of issues, and the issues in each stratum have similar propensity scores.

In propensity score matching, people usually rely on balance checks to determine if the matched samples (of both treated and untreated) are similar or comparable other than their treatment status. Specifically, we want to check if the covariate distribution is balanced between the treated and untreated samples.

Ideally, the issues within a stratum would be considered as matched, and their difference in outcome can be regarded as the treatment effect. However, if the propensity score estimation is not correctly specified, these issues may have quite different covariate distributions and may still not be comparable even if they share a similar propensity score.

Instead of significance testing, literature has suggested standardized mean difference (SMD) for assessing covariate balance [28], which is defined as:

$$SMD = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{(S_1^2 + S_2^2)/2}}, \tag{3.1}$$

where $\overline{X_1}$ and $\overline{X_2}$ are sample mean for the treated and control groups, respectively; S_1^2 and S_2^2 are sample variance for the treated and control groups.

Similarly, SMD for binary variable is defined as:

$$SMD = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)\right)/2}}, \tag{3.2}$$

where $\hat{p}_1$ and $\hat{p}_2$ are observed probability of the binary variables in the treated and control groups, respectively. There isn't a clear cut-off for SMD score. Some literature has suggested using 0.1 or 0.25 as reasonable cutoffs for acceptable standardized biases. [108].

We visualize the SMD score of all covariate-stratum combination as a 2-D heatmap in Figure 3.4. Such visualization allows us to assess the SMD score of each covariate in each stratum while also examine how the SMD score is distributed across covariates and strata. The left plot corresponds to the GBRT propensity score model, which we discussed earlier in this section. We see that most of the cells have a light color, indicating a good covariate balance. However, the two rightmost columns are much darker, with many cells of the darkest color. As discussed earlier, these columns correspond to the strata of extremely high propensity scores, and almost all issues in these strata use emoji. The dense dark spots indicate that these issues are poorly matched, which may suffer from a lack of common support. The rest of the heatmap still has scattered dark spots, but they are not as concentrated. To conclude, after trimming the 2 strata due to the lack of common support, we should be able to draw causal conclusions on the remaining 18 strata.

To check the sensitivity of the specification of our GBRT model, we trained another GBRT model with a slightly different set of features. Specifically, we exclude the main program language features in the new specification. Similarly, we plot the SMD distribution of all the covariates (including the main program language) in Figure 3.4 (middle). Similar to the correctly specified model, the misspecified model also yields a poor balance on the two strata of the highest propensity scores. The area of main program language covariates, which were left out in the misspecified model, is slightly darker than its neighbor, but the rest of the heatmap remains light with scattered dark spots, which is also similar to the correctly specified model. This sensitivity check confirms that the balance is relatively robust to misspecification.

We also plot the heatmap of SMD distribution based on a logistic regression model of the same feature set for comparing purpose (Figure 3.4, right). Aside from the two highest strata, whose imbalance is consistent with the GBRT models, we can see many more dark sports on the heatmap. The "Author Following," "Author Public Repos," "Body Length," and "Body Tokens" features are imbalanced in almost all propensity score strata. This is likely due to the non-linearity of these features and verifies the advantage of machine learning models in estimating propensity scores in complex scenarios.

3.5 Results

By matching issues with similar propensity scores, we assure that the distribution of the confounding factors is balanced within each stratum. In each stratum, issues with and without emojis are comparable, and the unbiased treatment effect (of using emojis) can be estimated. In Table 3.4, we report the test statistics of the dependent

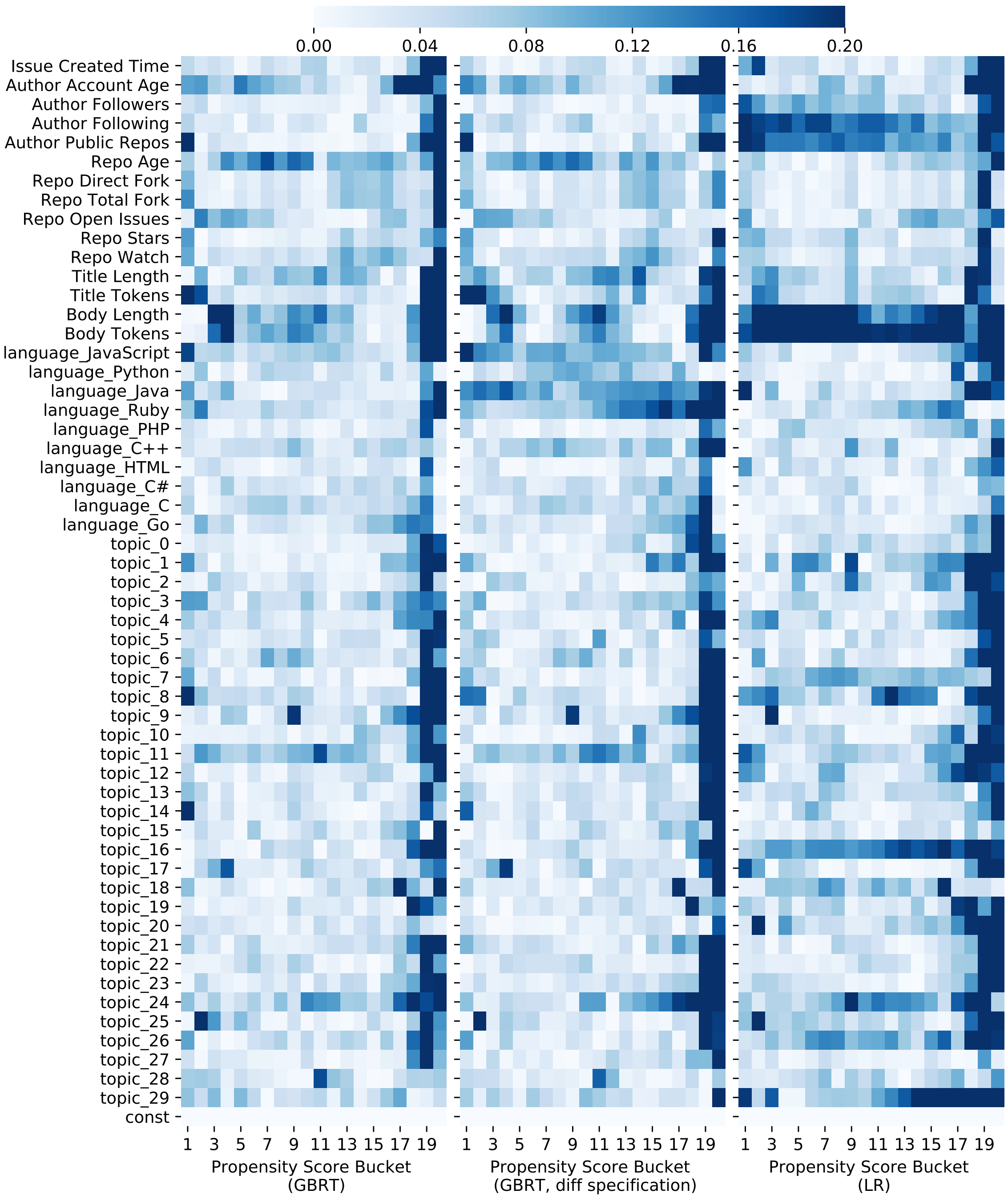


Figure 3.4: Balance Check for Propensit Score Estimation
The color of each cell represents the SMD score of one covariate (indicated on the y-axis) in the stratum (indicated on the x-axis). The darker the color, the larger the SMD score. We cap the maximum SMD at 0.2, and all SMD scores larger than 0.2 are also represented as the darkest color.

Table 3.4: The Effect of Using Emojis in Issues. Average Treatment Effect (ATE) and significance level (first column) are estimated by pooling treatment effects and variance in each stratum. We also report the observed difference and the average values in the second and third columns.

Hypothesis	Dependent Variable	ATE	Observed Δ	Avg.
H1	getting comment	0.054***	0.131	0.585
	# comments	0.27***	0.782	2.186
	# users who comment	0.161***	0.407	1.171
	# comments / user†	0.001	0.014	1.668
H2	prop. of issues closed in 30 days	0.017***	0.045	0.498
	issue closing time (days)‡	-0.370***	-0.633	4.357
H3	prop. of comments w/ emojis	0.110***	0.116	0.086

Significance level: *** 0.001, ** 0.01, or * 0.05 level.
† Computed on issues with one or more comment(s).
‡ Computed on issues closed in < 30 days only.

variables. For each outcome variable, we estimate the average treatment effect (ATE) by pooling the stratum-specific treatment effects [63]. And we pool the variances of the stratum-specific treatment effects as the estimate of variance [16, 85]. We calculate the z-score based on these estimates and report the significance level after adjusting for multiple hypotheses testing.

3.5.1 Increasing Participation

We first see if using emoji brings more discussions to an issue, which can be measured by the likelihood of getting comments or the number of comments. As shown in Table 3.4, issues with emoji are more likely to get comments (ATE = $5.4\%, p < 0.001$). And on average, issues with emoji get 0.27 more ($p < 0.001$) comments than those without. We may further decompose the effect into two factors, and check if this is because more users participate in the conversation or if the intensity of the participation increases. The former can be viewed as the extensive margin, while the latter as the intensive margin. We see that an issue with emoji attracts 0.161 more ($p < 0.001$) users to comment, however, for those who do post comments, their average number of comments under the same issues does not differ much between the treated and untreated. This significance along the extensive margin suggests that using emojis does attract more users' attention to join the discussion, while the insignificance along the intensive margin indicates that the activity level among participants is not affected.

Such results can be explained by the roles non-verbal cues play in Computer-Mediated Communication (CMC) [48, 49]. In the absence of facial expressions, non-verbal cues like emojis can express humor, or adjust the tone to be more friendly or less serious. (Consider the use of 🐛 for bugs!) With emoji, communication is made more funny and engaging, which attracts more participation from the audience. Note that the insignificance along the intensive margin is not surprising, and coincides with several other factors, including monetary incentives, which were initially believed to increase participation. For example, both [66] and [119] suggest increased monetary incentives draw more participation, but not necessarily of higher quality.

3.5.2 Resolving Issues

We have shown that emoji attract more discussions to issues. However, people do not simply want their issues to be watched; they want their issues to be resolved – bugs fixed, features added, questions answered. Does the increased participation brought by emoji actually help to resolve the issues? On GitHub, we can test this hypothesis (H2.1) by looking at the closing status of the issues. Specifically, an issue being closed usually indicates that the issue has been properly handled. We compare the proportion of issues being closed and the average closing time between those with and without emoji.

Indeed, we find that issues with emoji are more likely to get closed within 30 days ($\Delta = 1.7\%, p < 0.001$). Also, among those closed issues, the time spent before closing also decreases significantly for issues with emojis, with an average of 0.37 days

($p < 0.001$). Therefore, one can infer that the attention and participation the emoji attract are not from mere bystanders. Instead, the increased participation does help to resolve issues in a timely manner.

3.5.3 Reshaping Community Norms

With the help of propensity score matching, we have shown that emojis do increase participation, and such participation does help to resolve issues. Does the use of emoji have an effect beyond development tasks, but on community culture as well? Does emoji use by one user influence others? Are emoji reshaping GitHub community norms? Inspired by these questions, we test whether the use of emojis in an issue results in more emoji used.

Following the same process of propensity score matching, we test if using emoji in issues increases the use of emoji in their comments. From Table 3.4, we can see that more comments ($\Delta = 11.0\%, p < 0.001$) use emoji in reply to issues with emoji. There are two potential explanations for such an increase: emoji in the issues may raise awareness of emoji among the audience, and the audience may use emoji reciprocally to issues with emoji. This suggests that there is an upward spiral of using emoji. With such spiral, emoji are becoming the new norm of the GitHub community.

In Table 3.4, we also report the observed difference in the measured outcome between issues with and without emoji, without propensity score matching. In general, we see that the observed difference is several times larger than the estimated treatment effect. Such discrepancy evidences the need to adopt PSM in order to correct the selection bias.

3.6 Conclusion

With propensity score matching, we confirmed the effect of using emoji in increasing participation and resolving issues on GitHub. We also show that the use of emoji leads to more emoji usage in response, which is consistent with the increasing adoption of emoji. If we put this in the context of our end-to-end pipeline, the conclusion successfully identified using emoji as a potential nudge to promote the pro-social behaviors on the open-source platform.

This study draws several implications for open-source platforms and other online communities as well.

First, our work provides direct evidence of the positive effects of using emoji in user engagement and problem solving. In fact, the low ratio of posts containing emoji indicates a great opportunity for the GitHub community to promote emoji in conversations, through the designs of recommender systems or specialized interfaces. In general, one may expect that other visual features may have similar effects in engaging user participation. Narrowly speaking, GitHub and other developer communities like StackOverflow may consider adding more visual features to attract users into discussions, such as animations or GIF images. Broadly, adding visual designs into traditionally text-heavy tasks not only adds fun to the work, but may also help engage users in the tasks and even improve the quality of work.

Second, emoji may be an effective instrument for understanding and comparing different groups of users in online communities. On one hand, they are widely adopted in daily communications, yet they are not strongly tied to their different daily tasks, which makes them a suitable common ground to compare across different user groups. On the other hand, emoji are compact and usually have clear semantics, which eases the pain of natural language understanding for particular domains (e.g., dealing with cross-lingual texts, slang, professional vocabularies, or hashtags). Emoji are also associated with rich sentiment, which is convenient for analyzing the interpersonal relationships and emotional norms of online communities.

There are some limitations to our work. Although propensity score matching is employed to address the selection bias, it relies on the Conditional Independence

Assumption (CIA), which is hard to fully justify in our case. CIA assumes that X includes all confounding variables that affect the use of emoji and the potential outcome. However, some variables may not be observable or they may be hard to model. For example, project content and developer age may be unobserved confounders. The ideal way to estimate the effect of using emojis is to design a randomized experiment, which is beyond the scope of this empirical work.

Our paper only analyzes communication that happens through issues. However, communications may also go through other channels, such as Gitter, an instant-messaging service that connects easily with GitHub. This may have introduced biases to our analysis. Due to the lack of timestamps and user information, we did not include the emoji responses in our analysis and only focused on the emoji used in the free text. We may have underestimated the popularity of emoji and the proportion of users who used emoji.

A clear future direction is to study the heterogeneous effect of different emoji. For example, emojis of strong positive or negative sentiment may have different effects on the participation. It is also intriguing to conduct a finer-grained analysis, by classifying the issues into different purposes and classifying the users into different roles. Emojis may be used differently for different purposes and when the user takes certain roles in a collaborative project.

Finally, we cannot proceed with the second or third stages in our pipeline, as we are not able to implement an emoji recommender system on the GitHub platform. Nor can we test its effectiveness via field experiment. In the next two chapters, however, we implement and evaluate recommender systems in real-world settings, which allows us to examine the full potential of our pipeline.

CHAPTER IV

Recommending Teams Promotes Pro-social Lending in Online Microfinance.

In this chapter, we focus on the second and third stages of the end-to-end pipeline. We reports the results of a large-scale field experiment based on the hypothesis that group membership can increase participation and pro-social lending for an online crowdlending community, Kiva. The hypothesis was proposed in a companion study. The experiment uses variations on a simple email manipulation to encourage Kiva members to join a lending team, testing which types of team recommendation emails are most likely to get members to join teams as well as the subsequent impact on lending. We find that emails do increase the likelihood that a lender joins a team, and that joining a team increases lending in a short window (one week) following intervention. The impact on lending is large relative to median lender lifetime loans. We also find that lenders are more likely to join teams recommended based on location similarity rather than team status. Our results suggest team recommendation can be an effective behavioral mechanism to increase pro-social lending.

4.1 Introduction

Understanding strategies to increase pro-social behavior has important policy implications. Charities have explored various mechanisms to increase giving, such as seed money, matching gifts and peer pressure [10]. In comparison, an under-explored class of mechanisms utilizes group membership and inter-group competition [3, 110] to increase both participation and giving amounts. Compared to price-based strategies, such as matching gifts and rebates, empirical analysis of naturally-occurring data indicates that identity-based mechanisms have longer-lasting effects [32]. Our research explores two questions through a large-scale field experiment on a crowdlending community with a natural group structure (teams). First, which types of team recommendations are most likely to motivate lenders to join teams? Second, once they join a team, what is the subsequent impact on lending?

Our research is conducted at Kiva.org, a crowdlending community created to help micro and small enterprises in developing countries, which often lack access to the formal banking sector. Specifically, Kiva partners with local microfinance institutions to match individual lenders with low-income entrepreneurs in developing countries as well as selected cities within the United States. Through Kiva's platform, anyone can make a zero-interest loan of $25 or more to support an entrepreneur. Since its inception in 2005, Kiva has increased its membership significantly. However, while many lenders join Kiva for pro-social motives, they do not participate fully. Indeed, thirty-six percent of them have never made a single loan, and many others do not come back to Kiva after making their first loan [83]. Kiva's challenge is not unique, as many online contribution communities struggle with the issue of how to sustain member engagement and contributions.

To increase member engagement, some online communities have created group structures. For example, in 2008, Kiva instituted a lending teams program, a system through which lenders can create teams or join existing teams of other lenders. Once a team is created, it appears on Kiva's team leaderboard, which sorts teams by the total loan amounts designated to them by their team members. Since 2008, more than 38,957 Kiva teams have been created based on lender group affiliations such as organizations, geographic location, religious affiliation, or sports interests. Of note, many of the highly ranked teams are identity based, such as the "Atheists" and the "Kiva Christians." Each team has a dedicated forum where team members can coordinate their lending activities, ask and answer questions, and set goals for the team.

The use of groups to increase charitable contributions has intuitive appeal, but its success is difficult to measure with naturally-occurring field data because of sample selection bias. For example, lenders who join teams might simply be those who are more active in general [32]. To establish the *causal* relationship between group membership and pro-social lending, we use a randomized field experiment which enables us to combine the control of a laboratory experiment with the external validity of a field study [59, 35].

Our novel approach is inspired by the economic theory of social identity [3, 4] as well as the development of big data analytics in computer science. Research on social identity has consistently found that people derive their sense of identity from groups [36, 40]. This group identity can be used to increase voluntary contribution and improve coordination among team members in the laboratory [26, 51, 30, 42, 31, 29]. Building on these findings, we conduct a large-scale randomized field experiment to evaluate the effectiveness of team recommendation as a behavioral mechanism for increasing participation among Kiva members. Our approach enables us to synthesize the predictive accuracy of machine learning with the causal inference of economic theory and field experiments [72].

4.2 Literature Review

Our study builds upon findings from three streams of literature: charitable giving, advertising and recommender systems, and social identity. The charitable giving literature has uncovered several motivations and mechanisms for people to voluntarily give to charity [10]. In addition to the neoclassical preferences for public goods [23], people might derive a “warm glow” from the amount they give, which increases giving [6, 7]. People also respond positively to mechanisms which decrease the price of giving, such as tax subsidies [18], matching gifts or rebates [50, 70]. Sequential giving mechanisms [8, 86, 114], which utilize leadership gifts to transmit information or signal the value of the public good, have been shown to increase giving in the lab and field [96, 79]. Closely related to our study, researchers have shown both theoretically and experimentally that people might give because they care about their social image [9, 13], peer pressure [89], or social pressure [47]. In our context, when lenders join a team, team members can activate several of these mechanisms, such as leadership giving and social pressure, by posting messages on the team forum [32].

Our research is also related to the advertising literature. Recent field experiments show that advertising content, especially when it appeals to intuition, significantly affects demand [24]. More generally, personalized recommendations based on various machine learning algorithms have increased consumer adoption of recommended items, and have thus been widely used by e-commerce sites [98, 69]. Instead of recommending items, such as products, our study recommends lending teams to Kiva users.

Lastly, our study builds upon social identity theory [110, 3], and recent experimental research that uncovers the positive effects of group identity on voluntary contribution and coordination in the laboratory [26, 51, 30, 42, 31, 37, 29] and the field [53]. Our team recommendation approach extends social identity research to the realm of behavioral mechanism design at scale.

4.3 Experiment Design

4.3.1 Recommendation Algorithms

In our study, we use a lender's likelihood of joining a team to recommend teams based on both homophily and status. Homophily refers to the tendency to associate with similar others [88, 56]. As such, we recommend teams to lenders based on their similarity to the existing members of those teams. In our study, we use two different measures of homophily: location similarity and loan history similarity. The former is based on the number of lenders in a team who share the same location as the target lender, whereas the latter is based on how often the lenders have lent to the same borrowers. In addition to homophily, we recommend teams based on status [104], using the top three teams on the Kiva leaderboard as the high-status teams. The details of the recommendation algorithms are illustrated as follows.

Recommendations based on team status The simplest recommendation strategy is to recommend teams that are ranked highly on the team leaderboard. Kiva provides several leaderboards that rank teams based on either the total loan amount attributed to the team or the number of team members, in the most recent month or all time. For the experiment, we use the default leaderboard that lenders see when they visit the Kiva Team page, the all-time total amount lent.

Note that every lender receives the same recommendations under this strategy. The three teams we recommend to the lenders are "Atheists, Agnostics, Skeptics,...", "Kiva Christians," and "Guys holding fish."

Recommendations based on location similarity The goal of this algorithm is to recommend the most popular teams in a lender's local area. This is motivated by the fact that there are many location-based teams on Kiva and by the conclusion of our previous work that the maximum location similarity between a lender and all the teams is partially correlated with whether the lender has joined a team [32]. This also reflects the results of an online data mining competition we ran with doctoral students at the University of Michigan using the Kiva API data. The following algorithm, written by the first author, is the one that performed best in that competition. We calculate the location similarity between two lenders u and v as $l_{uv} \in \{0, 1, 2\}$ [32]. If the two lenders are from different countries, $l_{uv} = 0$. If two lenders are from the same city, $l_{uv} = 2$. The condition for $l_{uv} = 1$ includes the following two cases: 1) if the two lenders are not in the same city but in the same state in the United States or Australia, or the same province in Canada, or 2) if they are from the same country other than the United States, Australia or Canada. This is because there are significantly more lenders on Kiva from the United States, Australia or Canada than from any other country.

The location similarity of a team t in the neighborhood of a lender u is calculated as the sum of the location similarities between that lender and all lenders in that team. That is, $L(u, t) = \sum_{v \in T} l_{uv}$, where T denotes the set of lenders belonging to team t. For every lender, we rank all teams by the location similarity of these teams and recommend the three highest-ranked teams. For these recommendations, we exclude the three teams highest on the leaderboard: “Atheists, Agnostics, Skeptics,...,” “Kiva Christians,” and “Guys holding fish,” for two reasons. First, the Atheists and Christians are outliers in that they overwhelm all other teams in size. Consequently, they often appear as winners of location-similarity based recommendations. Second, to differentiate between status-based and homophily-based recommendations, we exclude all three teams.

Recommendations based on loan history similarity We also construct a recommender system based on the loan history of a lender. This is motivated by the homophily conjecture that lenders who lend to similar borrowers share similar interests and are thus more likely to join the same teams.

Borrowers on Kiva are registered in 80 countries from 8 geographical regions (Oceania, Asia, etc). They loan to facilitate 149 types of activities which are further categorized into 15 sectors. Let $\mathcal{S}_u$ be a set of loans made by a user u and $\mathcal{S}_t$ be a set of loans that are attributed to a team t. The **relevance** of the team to the user is scored by the following function:

$$Relevance(u, t) = \sum_{i \in \mathcal{S}_u} \sum_{j \in \mathcal{S}_t} [f_g(i, j) + f_a(i, j)], \tag{4.1}$$

where $f_g(i, j)$ equals 2 if the two loans i and j are from the same country, 1 if they are from two different countries in the same region, and 0 if they are not from the same region; $f_a(i, j)$ equals 2 if the two loans i and j are for the same activities, 1 if they are for different activities in the same sector, and 0 if they are not for activities in the same sector.

Note that the relevance score as defined in Equation (4.1) favors large teams that have made many loans. We further normalize the score by taking into account the total number of loans made by each team. That is:

$$Normalized\_Relevance(u, t) = \frac{Relevance(u, t)}{|\mathcal{S}_t| + 50}. \tag{4.2}$$

Given a user who has not joined a team, we calculate the normalized relevance score for every team and recommend the three top-scoring teams to that user. For consistency with the recommendations based on location similarity, we also exclude the top three teams on the leaderboard, “Atheists, Agnostics, Skeptics,...,” “Kiva Christians,” and “Guys holding fish,” for these recommendations.

Table 4.1: Summary of Experimental Treatments.

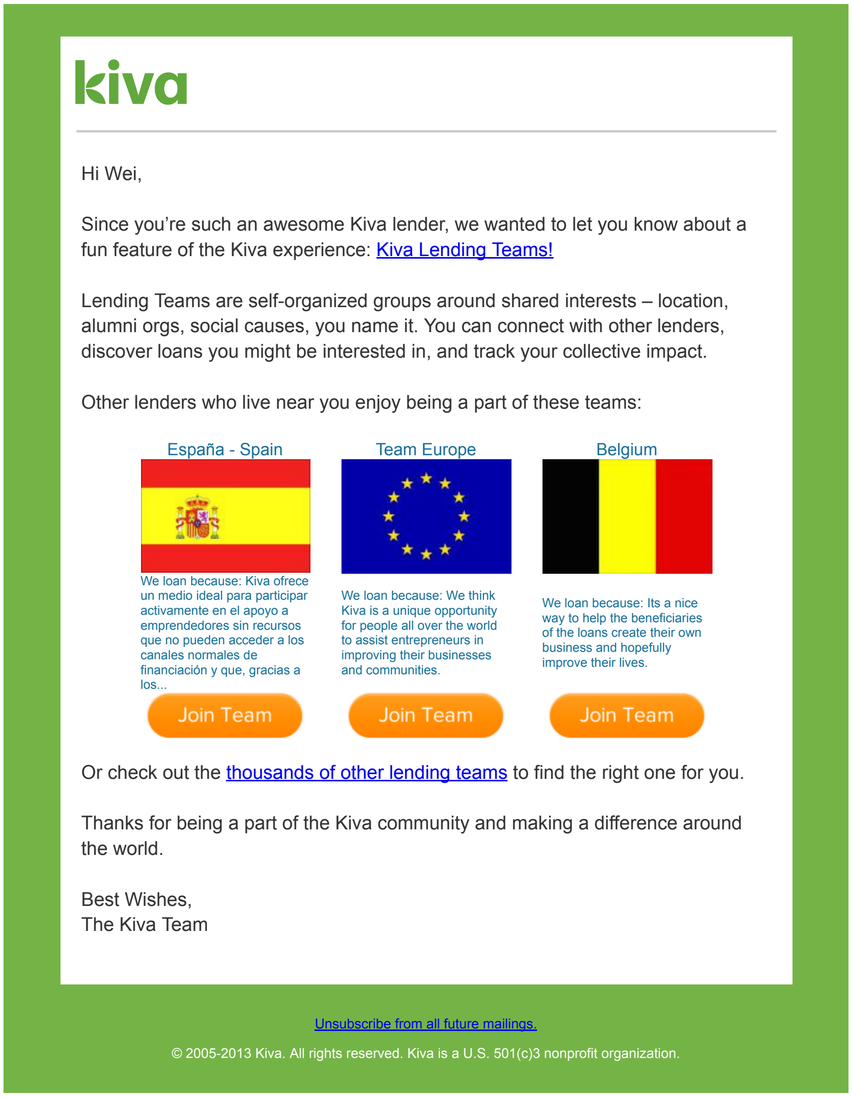
		Explanation of Recommender Algorithm	
		Explanation	No Explanation
Recommendation Algorithm	Location	Location-Explanation	Location-NoExplanation
	Loan History	History-Explanation	History-NoExplanation
	Leaderboard	Leaderboard-Explanation	Leaderboard-NoExplanation
Control		No Contact	
Placebo		Teams Exist	

4.3.2 Factorial Design

We employ a 3×2 factorial design (Table 4.1). Along one factor, we vary our recommendation algorithms along one factor based on lender-team location similarity, loan history similarity, or team status.Along the other factor, we vary whether our recommendation rationale is explained to the lender. Literature suggests that providing an explanation can increase the acceptance of a recommendation [60, 98]. By varying whether a lender receives an explanation, we can obtain a better understanding of whether a factor impacts the effectiveness of the recommender system. We also include a control condition where we do not contact lenders (no contact) and a placebo condition where we email lenders to make them aware that there are lending teams on Kiva without providing any specific recommendations (teams exist) to control for any contact effect. The text of the email is completely identical across treatments, except for the variables that change across treatments. Figure 4.3.2 presents a sample email from the Location-Explanation treatment.

Each email consists of three parts. Part 1 is common to all treatments and the placebo,

"*Hi [FirstName], Since you're such an awesome Kiva lender, we wanted to let you know about a fun feature of the Kiva experience: Kiva Lending Teams! Lending Teams are self-organized groups around shared interests – location, alumni orgs, social causes, you name it. You can connect with other lenders, discover loans you might be*



Hi Wei,

Since you're such an awesome Kiva lender, we wanted to let you know about a fun feature of the Kiva experience: Kiva Lending Teams!

Lending Teams are self-organized groups around shared interests – location, alumni orgs, social causes, you name it. You can connect with other lenders, discover loans you might be interested in, and track your collective impact.

Other lenders who live near you enjoy being a part of these teams:

España - Spain

We loan because: Kiva ofrece un medio ideal para participar activamente en el apoyo a emprendedores sin recursos que no pueden acceder a los canales normales de financiación y que, gracias a los...

Team Europe

We loan because: We think Kiva is a unique opportunity for people all over the world to assist entrepreneurs in improving their businesses and communities.

Belgium

We loan because: Its a nice way to help the beneficiaries of the loans create their own business and hopefully improve their lives.

Or check out the thousands of other lending teams to find the right one for you.

Thanks for being a part of the Kiva community and making a difference around the world.

Best Wishes,
The Kiva Team

Unsubscribe from all future mailings.

Figure 4.1: An Email Screenshot of the Location-Explanation treatment.

interested in, and track your collective impact."

Likewise, each email ends with Part 3,

"*[Or] Check out the thousands of [other] lending teams to find the right one for you.*"

"*Thanks for being a part of the Kiva community and making a difference around the world.*"

While the text of emails sent to lenders in the placebo ("teams exist") condition consists of Parts 1 and 3, lenders in the six treatments also received one of the following in the second part of the email:

1. Leaderboard with explanation treatment (Leaderboard-Explanation):

   "*Some of the most popular teams are: [TEAMS].*"

2. Location similarity with explanation treatment (Location-Explanation):

   "*Other lenders who live near you enjoy being a part of these teams: [TEAMS].*"

3. Loan history similarity with explanation treatment (History-Explanation):

   "*Based on your past lending, people who have made similar loans enjoy being a part of these teams: [TEAMS].*

4. Recommendations without explanations treatments (Leaderboard-NoExplanation, Location-NoExplanation, History-NoExplanation)

   "*Here are a few teams you may want to check out: [TEAMS].*"

4.3.3 Experimental Procedure

The experiment is conducted in 2014. We use a group of 69,845 lenders who have made at least two loans in the past six months but have never joined a team. We

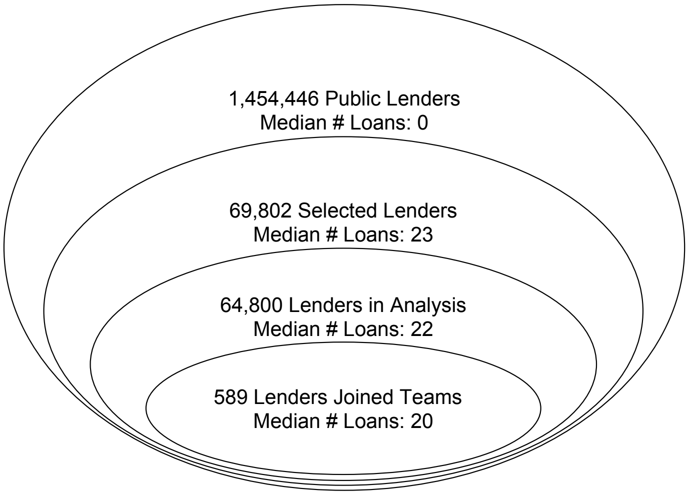


Figure 4.2: Sample and Population Comparison.

The number of lenders and median number of loans of all public users, those who are selected as participants, those whose data is used in our analyses, and those who joined at least one team during our experiment.

then randomly assign each lender to one of eight experimental conditions with equal probability.[1]

We then assign each user to one of the treatments, the placebo, or the control condition using stratified randomization. The stratified random assignment is based on the total loan amount by each lender before the experiment. We want to ensure that the most active Kiva lenders are not all concentrated into one treatment, so we rank the lenders based their total loan amounts, taking the top 8 lenders and randomly assigning them to different conditions. We then repeat this for each group of 8 lenders, proceeding down the ranked list. Between assigning lenders to conditions and running the experiment, 43 users joined a team and were dropped from our sample. This yields a final sample of 69,802 users. The size of the sample and population is summarized with a Venn Diagram in Figure 4.2.

Before running the experiment, we run pair-wise Kolmogorov-Smirnov tests of the

[1] Based on Kiva Privacy Policy and the information need of our recommendation algorithms, we include only lenders that set their pages and loans to public in their account settings, allow marketing emails in their communication settings, and provide location information in their profile.

Table 4.2: Lending Statistics of Each Treatment during 6 months prior to experiment.

Experimental Condition	# of Users	Lending Statistics (average)			
		Amount Loaned	# Loans	Repayment Term	Account Balance
No-Contact	8725	184.29	6.07	18.50	36.24
Teams-Exist	8725	181.15	5.96	18.33	35.89
Location-Explanation	8726	181.34	6.04	18.45	35.22
Location-NoExplanation	8726	182.68	6.02	18.32	37.13
History-Explanation	8726	181.54	5.93	18.29	37.89
History-NoExplanation	8725	181.78	5.94	18.38	35.62
Leaderboard-Explanation	8723	182.14	6.05	18.40	34.37
Leaderboard-NoExplanation	8726	195.83	6.51	18.28	37.89

Note: Pairwise Kolmogorov-Smirnov tests comparing each experimental condition with the other yield $p > 0.10$ for each observable characteristic. Amount Loaned and Account Balance are in United States dollars, whereas Repayment Term is in months.

equality of distributions based on the user statistics to verify that our randomization produces balanced treatments across observable characteristics. The results of these tests show that the number of loans, average amount per loan, balance, average loan terms for fundraising or repayment, and auto-lending settings do not differ significantly at the 10% level between any treatments. Thus, the Kolmogorov-Smirnov tests do not reject the hypothesis that these values are drawn from the same distribution. We summarize the lending and location statistics of each treatment in Table 4.2.

We send each lender in our treatment groups an email from Kiva, for a total of 61,077 emails. After excluding lenders whose emails bounced and those who made their accounts private, we have a total of 64,800 lenders whom we intend to treat (henceforth *All*). Of these lenders, we find that one-third ($n = 20,371$) open our email, constituting our treated sub-sample (henceforth *Opened*). We then track the team-joining and lending behavior of each lender for the next two months. Anonymized data will be available from the open ICPSR data repository. Our research protocol was approved by the University of Michigan IRB (HUM00050208), which exempted us from obtaining informed consent.

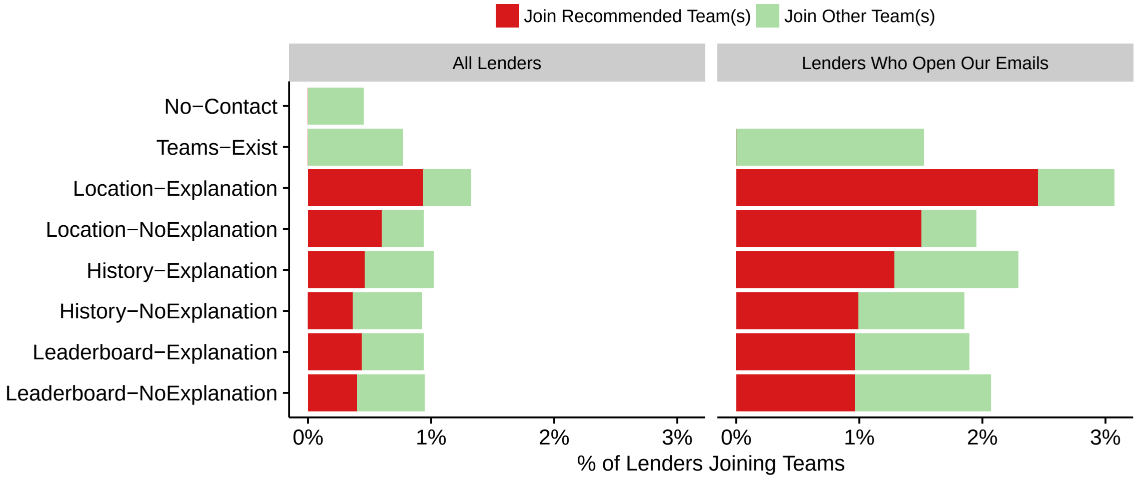


Figure 4.3: Proportion of Lenders Joining Teams in each Experimental Condition. This figure presents the proportion of lenders who join a lending team in each experimental condition after our email intervention. Location-based recommendations exhibit a higher proportion of lenders joining recommended teams (67.96%), compared to lending history similarity (42.31%) or leaderboard-based (44.37%) recommendations ($p < 0.01$, proportion of t-tests). Similar results are observed when we focus on lenders who open our email (right panel).

4.4 Results

We first examine what types of recommendations are most effective in increasing team membership. Figure 4.3 presents the proportion of lenders who join a lending team in each treatment after our email intervention, for both all lenders (left panel) and those who open our emails (right panel). For both groups, lenders who receive a location similarity explanation are most likely to join a team, accounting for 3% of the group who open their emails. This participation rate is comparable to that in other charitable-giving field experiments using mailing campaigns [79, 70].

We next conduct a regression analysis (Table 4.3 and Figure 4.4) and find that every treatment leads to a significantly higher likelihood of joining a team, compared to the no-contact control condition, for both the all lenders (column 1) and opened-

Table 4.3: Treatment Effects on the Likelihood of Joining Teams: Probit Regressions. Marginal effects reported, calculated at the mean level of the covariates. (a) The decision to join a team is regressed on the seven treatment dummies for all lenders in our sample ($n = 64,800$). (b) The second model uses the same specifications but is restricted to the lenders who opened their emails or were not contacted ($n = 29,055$). (c) The third model is restricted to lenders who were sent emails and opened them ($n = 20,371$). Applying a multiple hypothesis testing correction [80] yields the same significance levels as above, except for the "History-Explanation" variable in column (3) which becomes insignificant at the 10% level.

Dependent Variable: Joined a team			
	(1)	(2)	(3)
	All Users	Opened No-Contact	Opened
Team-Exist	0.0045***	0.0155***	
	(0.002)	(0.003)	
Location-Explanation	0.0094***	0.0256***	0.0145***
	(0.002)	(0.002)	(0.004)
Location-NoExplanation	0.0062***	0.0189***	0.0050
	(0.002)	(0.002)	(0.004)
History-Explanation	0.0070***	0.0212***	0.0083**
	(0.002)	(0.002)	(0.004)
History-NoExplanation	0.0061***	0.0182***	0.0039
	(0.002)	(0.003)	(0.004)
Leaderboard-Explanation	0.0062***	0.0185***	0.0043
	(0.002)	(0.002)	(0.004)
Leaderboard-NoExplanation	0.0063***	0.0197***	0.0062
	(0.002)	(0.002)	(0.004)
Number of Subjects	64,800	29,055	20,371

1) Standard errors in parentheses.

2) Significant at the: * 10%, ** 5%, and *** 1% levels.

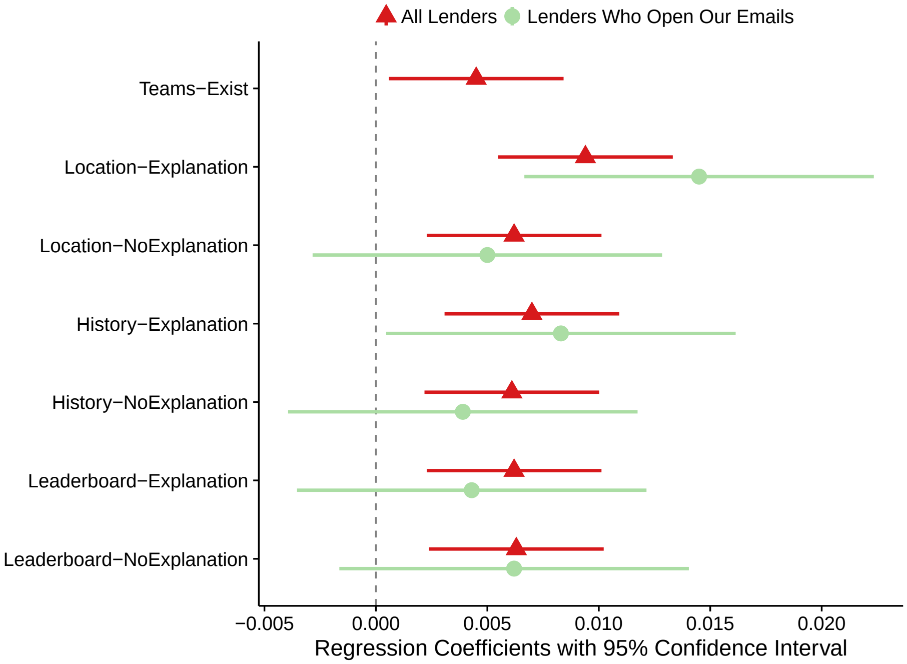


Figure 4.4: Treatment effects on the likelihood of joining teams. This figure presents the treatment effects on the likelihood that a lender joins a lending team (Table 4.3). When we focus on all lenders (lines with red triangle), we find that every treatment significantly increases the likelihood of joining a team compared to the control condition. When focusing on lenders who open our email (lines with green circle), we find that the homophily-based recommendations with an explanation also significantly increase the likelihood of joining a team, compared to the teams-exist condition. Explanations increase the likelihood of joining a team for only the location-based recommendations (All: $p = 0.02$; Lenders who open our email: $p = 0.01$; Wald tests).

email (column 2) groups ($p < 0.01$). Of those who open their emails, lenders in the location similarity with explanations treatment are more likely to join a team compared to those in the teams-exist condition ($p < 0.01$). These results are robust to a multiple hypothesis testing correction [80].

We next explore the types of team lenders most likely to join by examining the characteristics of teams joined by our lenders. Table 4.4 displays the results of eight conditional logit specifications with odds ratios reported, with one specification per treatment. In our regressions, we use whether each lender joined each team as our dependent variable, and location similarity, loan history similarity, team status, team size, and experimenter recommendation as our independent variables.

The results for our control and teams-exist conditions (columns 1 and 2) show that lenders are more likely to join teams with higher location similarity and status. The odds of a lender joining a team whose location similarity is 1 percentile higher is 2% higher, while the odds of a lender joining a top ten team is 13 times higher than those of joining a non-top ten team. On the other hand, we find that neither lending history nor team size impacts lenders' choices. These findings show that lenders value both homophily and status when deciding to join a team. It is also noteworthy that location and status information are easily found on Kiva's website while lending histories are more difficult to locate.

Interestingly, we find that the provision of a location similarity recommendation mitigates the influence of team status, leading lenders to join recommended teams or teams with higher history similarity (columns 3 and 4). By contrast, our recommendations based on loan history similarity (columns 5 and 6) do not substantially change how lenders choose their teams. Finally, recommendations based on team status (columns 7 and 8) seem to change lender behavior in a way similar to that of location-based recommendations, but only when we explain our recommendations.

Finally, we study whether joining a team increases pro-social lending. To address

Table 4.4: Choice Model: Conditional Logit Regressions.

	Dependent Variable: Joined a Team							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	No-Contact	Team-Exist	Loc.-Exp	Loc.-NoExp	Hist.-Exp	Hist.-NoExp	Lead.-Exp	Lead.-NoExp
Location Similarity (Percentile)	1.03***	1.02***	1.02***	1.05***	1.02***	1.01***	1.02***	1.02***
	(0.011)	(0.005)	(0.006)	(0.017)	(0.005)	(0.005)	(0.007)	(0.005)
History Similarity (Percentile)	1.00	1.00	1.02**	1.01	1.01	1.00	1.02**	1.00
	(0.007)	(0.007)	(0.008)	(0.009)	(0.011)	(0.008)	(0.010)	(0.010)
Top Ten Team	13.07***	13.60***	0.81	1.01	6.98***	13.85***	16.74***	6.22***
	(6.476)	(5.305)	(0.264)	(0.390)	(2.759)	(5.662)	(8.326)	(2.535)
Team Size (Percentile)	1.00	1.00	0.99	0.99	1.00	1.01*	0.98***	1.01
	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)	(0.008)	(0.008)	(0.012)
Recommended			82.39***	37.32***	119.52***	213.82***	7.78***	7.26***
		(27.780)	(14.638)	(36.696)	(71.360)	(2.473)	(2.504)	
Number of Teams	491	491	491	491	491	491	491	491
Number of Subjects	35	61	105	74	80	72	72	74

1) Standard errors in parentheses, clustered at the subject level.

2) Significant at the: * 10%, ** 5%, and *** 1% levels.

Odds ratios reported. Whether the subjects join teams is regressed against the two similarity measures (coded as the percentile of the measure for each subject-team pair), whether the team is one of the top teams, the team size, and whether or not the team was recommended through the experiment. This regression is performed separately for each treatment. While team size never significantly affects the joining decision, and a recommendation always significantly increases the likelihood of joining a team, the effects of the other variables depend on the treatment. When the teams are recommended based on either lending history (columns 5 and 6) or the leaderboard (columns 7 and 8), both similarity measures and whether the team is a top ten team significantly increases the likelihood that the subject joins the team. A location recommendation (columns 3 and 4) causes subjects to ignore the top ten teams. Compared to the cases where no recommendation is made (columns 1 and 2), any type of recommendation increases the degree to which subjects pay attention to lending history. When no recommendation is made, lending history similarity decreases subjects' likelihood of joining a team.

Table 4.5: Difference-in-Differences Regressions of Average Daily Lending Amount (2SLS).

	1st Stage	2nd Stage: Average Amount			OLS		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		1-Day	7-Day	30-Day	1-Day	7-Day	30-Day
Email	0.0053***						
	(0.001)						
Join Team		298.558***	55.914***	10.231	5.257***	0.566*	0.517***
		(72.283)	(21.058)	(7.318)	(0.755)	(0.337)	(0.134)
Constant	0.0045***	-2.660***	-0.936***	-0.236***	0.007	-0.433***	-0.147***
	(0.001)	(0.670)	(0.195)	(0.068)	(0.072)	(0.032)	(0.013)
Obs.	64,800	64,800	64,800	64,800	64,800	64,800	64,800

1) Standard errors in parentheses.

2) Significant at the: * 10%, ** 5%, and *** 1% levels.

The endogenous variable, whether a lender joins a team ("Join Team"), is instrumented with whether a lender receives an email in the experiment ("Email"). As the results of a two-stage least squares instrumental variable regression, the coefficients on the "Join Team" variable in columns (2)-(4) give local average treatment effects, or the effects on the subset of lenders who only join a team because of our email ("compliers"). The different columns give different window sizes in a difference-in-differences setting. The effect is significant up to a week after we send the email. Ordinary least squares estimates are also displayed in columns (5)-(7) for comparison. The difference between the IV and OLS estimates is due to the difference in the local average treatment effects (given by the IV regressions), which only gives the effect on compliers, and average treatment effects (given by the OLS estimates, though with potential selection bias), which gives the effect on all subjects. There are a large number of lenders who do not join any team in our sample, and the effects of our treatment on these subjects are not captured by the IV estimates.

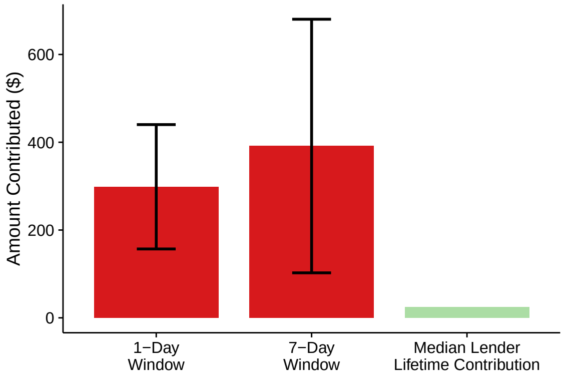


Figure 4.5: Effects of team membership on pro-social lending. This figure reports the results of our two-stage least squares instrumental variable regression coefficients (Table 4.5), indicating the effects of joining a lending team on contributions for the 1-day (left red bar) and 7-day (middle red bar) window. The median Kiva lender's lifetime contributions ($25) is plotted to provide a benchmark (green bar).

any potential endogeneity issues caused by self selection, we use the random treatment assignment in our experiment, namely whether the lender received an email, as an instrumental variable for joining a team. Figure 4.5 and Table 4.5 display the results of our two-stage least squares instrumental variable regression. In the first stage, we find that the "Email" variable, denoting whether a lender received an email, is not a weak instrument for joining a team, with an F-statistic of 23.55. Next, for this instrument to satisfy the exclusion restriction, it must be the case that an email does not directly affect lending except through increasing the likelihood that a lender joins a team. This might occur if contacting the lenders regarding Kiva reminds them of Kiva's existence, prompting them to lend. However, since our previous field experiment on Kiva has shown that simply contacting the lenders does not affect lending [32], we conclude that the instrument satisfies the exclusion restriction.

This regression employs a difference-in-differences approach. For three different window sizes, the dependent variable in each second-stage regression is the difference in total loan amounts t days before and after our treatment, where t is the window size. Thus, the coefficients on the "Join Team" variable indicate how much more lenders who join teams give than those who do not join teams after the treatment, controlling for the same difference before the treatment. The results of this regression show that joining a team significantly increases lending. However, it is important to note that since these estimates are derived from an instrumental variables regression, they give the local average treatment effect, not the average treatment effect [65]. Therefore, the estimates apply only to lenders who would join a team if prompted by an email.

This effect is also insignificant beyond one week. One possible reason for the lack of an observed long-term effect is that lenders may wait until initial loans are repaid before lending again, a process which may take 12-18 months. However, even the one-week effect ($392) is more than fifteen times the lifetime contribution of the median Kiva lender ($25), indicating that team membership is effective in increasing member contributions on those lenders who would join a team because of our email.

4.5 Discussion

This paper reports the results of a large-scale field experiment designed to test the hypothesis that team membership can increase participation and lending for an online crowdlending community, Kiva. We find that emails increase the likelihood that a lender joins a team, and that joining a team increases lending in an one-week window following the decision to join. While this experiment does not explore the mechanism through which joining a team increases giving, our prior empirical analyses and field experiment point to two mechanisms at work [32]. First, joining a team increases information sharing about specific borrowers on the team forum, which reduces team members' search costs and increases their lending. Second, joining a team increases the pressure to help improve the team's ranking on the Kiva leaderboard. Therefore, effective teams share information and coordinate their loans to reduce search costs, and emphasize team competition through goal setting. Our results suggest that recommending teams to members of an online lending community based on homophily is an effective mechanism to engage community members and increase their contributions.

CHAPTER V

Putting Organization into the Gig Economy: A Field Experiment at a Ride-sharing Platform

The gig economy provides workers with the benefits of autonomy and flexibility, but it does so at the expense of work identity and co-worker bonds. These sacrifices make gig workers less productive and more likely to leave. In this study, we examine the effect of team formation on the productivity of drivers at a ride-sharing platform. Specifically, we use social identity theory to develop a team formation and inter-team contest field experiment at DiDi, the dominant ride-sharing platform in China. In our study, we assign drivers to teams either randomly or based on homophily in age, hometown location, or productivity, and we have these teams compete for cash prizes. Our results show that platform designers can leverage team identity to increase productivity in a gig economy, especially when teams are formed to facilitate member communication.

Compared with the previous two chapters, we do not have the empirical data to support causal discoveries. Nor do we have behavior data to build a recommender system with machine learning algorithms. Instead, we start directly from the third stage and conduct a large-scale field experiment, which confirms the flexibility of our proposed end-to-end pipeline.

5.1 Introduction

As trends in work sourcing move us toward a gig economy, this economy is widely considered to be the future face of work, despite questions about its sustainability. While workers in traditional sectors derive their identities from their work and share their experiences with co-workers, those whose livelihood relies on the gig economy often find that "these are jobs that don't lead to anything," citing a lack of work identity and bonds with co-workers as well as an inability to move upward based on strong performance (*The New Yorker*, May 15, 2017).

To analyze these and other concerns associated with the gig economy, we apply social identity theory [3] to a large online platform, Didi Chuxing (DiDi henchforce), where individual drivers offer ride sharing in China. Specifically, we design a field experiment to study team formation and inter-team competition within DiDi. In our experiment, we examine how the creation of an organization identity impacts driver productivity. Furthermore, since DiDi is a flat organization with no group structure, we are also able to investigate how different team formations impact team member communication and productivity.

Our research applies insights from identity economics [3, 4]. This research shows that, when people feel a stronger sense of common identity with a group, they exert more effort and make more contributions to public goods to reach a more efficient outcome [51, 31]. Applying this theoretical framework to our setting, we anticipate that a driver who has a strong sense of team identity will work harder to help his team get ahead compared to drivers who do not belong to any team.

In examing how different team formations may have different effects on communication and coordination, we use an algorithm that maximizes either similarity or diversity within a team. We conjecture that similarity might facilitate team member communication and coordination, leading to intra-team bonding and team stability [103, 120, 71]. Indeed, empirical network science studies provide evidence for homophily, or the tendency of people to associate with others whom they perceive as similar to themselves in some way [88, 56]. By contrast, we conjecture that diversity might bolster team performance, due to different perspectives in problem-solving and better complementarity among team members [74].

In addition to examining different team formation strategies, we draw on insights from contest theory to explore how team identities form [73]. In our experiment, we apply a theoretical model of team contests with multiple pairwise battles by having subjects engage in inter-team contests for cash prizes, which have been shown to be among the most effective ways to strengthen team identity [51, 54].

Lastly, our work contributes to the rapidly growing literature on the ride-sharing economy, which has uncovered important insights related to labor market outcomes [58], consumer surplus [39], and decentralized dynamic matching efficiency [81]. Our findings contribute to this stream of research by showing that a team-based approach can significantly improve driver productivity.

5.2 Experiment Design

To test the effectiveness of team formation and inter-team competition on productivity, we design a multistage natural field experiment using the ride-sharing platform DiDi.

Recruitment stage: We conduct our experiment in the southern city of Dongguan, China. We begin with 480,000 DiDi drivers registered in the city of Dongguan. We first apply a set of eligibility criteria for participation in our study to satisfy a minimum threshold of activities in the two weeks prior to the start of the experiment, yielding 29,384 eligible drivers. From this group, we randomly choose 24,000 to invite to participate in our experiment. From the invited group, 2,343 drivers accept our invitation, with 531 of these indicating interest in being a team captain. We then randomly place our invitation respondents into five treatment and one control condition, each consisting of 350 drivers. The remaining 283 drivers serve as backups in case drivers in the treatments drop out before the start of the contest (for details, see Section 5.5).

Team formation stage: In each treatment condition, we partition the 350 drivers into teams of 7. Based on findings on team formation from previous studies [1], we select five dimensions by which to form our teams: hometown similarity, age similarity, productivity similarity, productivity diversity, and random formation.

Our first dimension, hometown similarity, is based on previous findings that location similarity is the most effective characteristic in getting a microfinance community member to join a specific lending team [1]. In our study, we use hometown similarity, a form of location similarity, assigning drivers from the same (or a nearby) province to the same team. Our second dimension, age similarity, is based on prior research illustrating the importance of good communication for teams to be sustainable [32]. We conjecture that people of a similar age might find it easier to communicate and thus form our age-similarity teams to reflect an age span of 5-10 years. Third, we include productivity similarity as one of our strategies as it is the preferred team formation strategy by the platform. Finally, we draw on recent scholarly research supporting the advantages of diversity [95]. and use two strategies to create diverse teams. To achieve productivity diversity in our teams, we partition drivers into seven buckets based on their productivity in the two weeks prior to the announcement of the team contest. Each team consists of drivers from all seven buckets. Our final strategy, random formation, reflects the diversity achieved from a random grouping of drivers. Details of our team formation algorithms are relegated to SI. In sum, our team formation strategy yields a total of 1,750 treatment drivers formed into 250 teams, with 50 teams in each treatment.

Within each team, we identify a team captain who is notified of this position, given the phone number of each team member, and asked to complete a survey. The

Table 5.1: Prize Structure.

Prize Structure	Individual Win	Team Win
Individual-Prize Treatment	30	-
Team-Prize Treatment	-	30
Hybrid-Prize Treatment	15	15

This table indicates the prize that drivers get if they win the individual contests (individual win), if their teams win a majority of the contests (team win), or both. The prize is calculated for each contest based on the number of trips a driver makes on that day.

survey requires captains to communicate with each driver in the team to get their license plate numbers as well as several key pieces of demographic information (see SI for the pre-contest survey questions and summary statistics). Meanwhile, team members are given the captain's phone number and told that the captain might call them. The initial team task is designed in such a way as to nudge the captains to initiate communication with their team members. Captains who fill in the survey through an online form are given 100 CNY as a bonus regardless of the correctness of their answers. If a captain submits the survey, we mark the team as *responsive*. In our sample, 60.8% of our captains submit their survey.

Contest stage: Our contest rules are based on a theoretical model of team contest [54]. Results are determined by multiple pairwise battles. Specifically, we set up a contest where drivers from two rival teams form pairwise matches to fight distinct component battles. In this contest, a team wins if and only if its players win a majority of their battles. In the theoretical model, each driver receives a private reward from winning her own battle as well as the benefits from their team's winning. Under these contest rules, we obtain the desirable neutrality results, that is, the outcome is history, sequence, and temporal-structure independent.

In our experiment, we decompose the effects into individual, team, and hybrid prize allocation conditions, as illustrated in Table 5.1. Under the individual prize condition, the driver who wins the contest receives a 30 CNY prize, regardless of team performance. Under the team prize condition, each driver in a team that wins a majority (4 or more) of its contests receives a 30 CNY prize. Under the hybrid prize condition, drivers receive both individual and team prizes of 15 CNY each. The prizes are set such that the expected reward per driver remains the same across treatments, which is 15 yuan under the symmetry assumption.

In our ride-sharing context, since we are conducting a field experiment, our drivers also earn piece rate in addition to any prize money. This differs from the theoretical model, where incentives come solely from prizes. To determine our pairwise matching, we sort the 250 teams decreasingly by productivity (the sum of the individual productivity of team members) in the two weeks prior to the announcement of the contest. From the groupings of most to least productive teams, two adjacent teams are paired for each contest, independent of their team formation strategy. This matching process ensures that each pair of teams in each contest is as similar as possible, preserving the symmetry assumption from the theoretical model. We randomly assign each team-pair into one of three prize allocation conditions with equal probability. Finally, within each team, we use an algorithm to automatically pair drivers by their productivity, i.e., the most productive driver in team A competes with the most productive one in team B, and so on. The drivers compete on the number of trips they finish in one day of competition.

The contest was implemented between August 13 - 21, 2017, with one day off between every two contest days. Before each contest day, we reset the contest and repeat it five times with the same pairing of teams. The contest results are calculated at the end of each contest day and communicated to each driver on the following day. Figure 5.1 describes the experimental process.

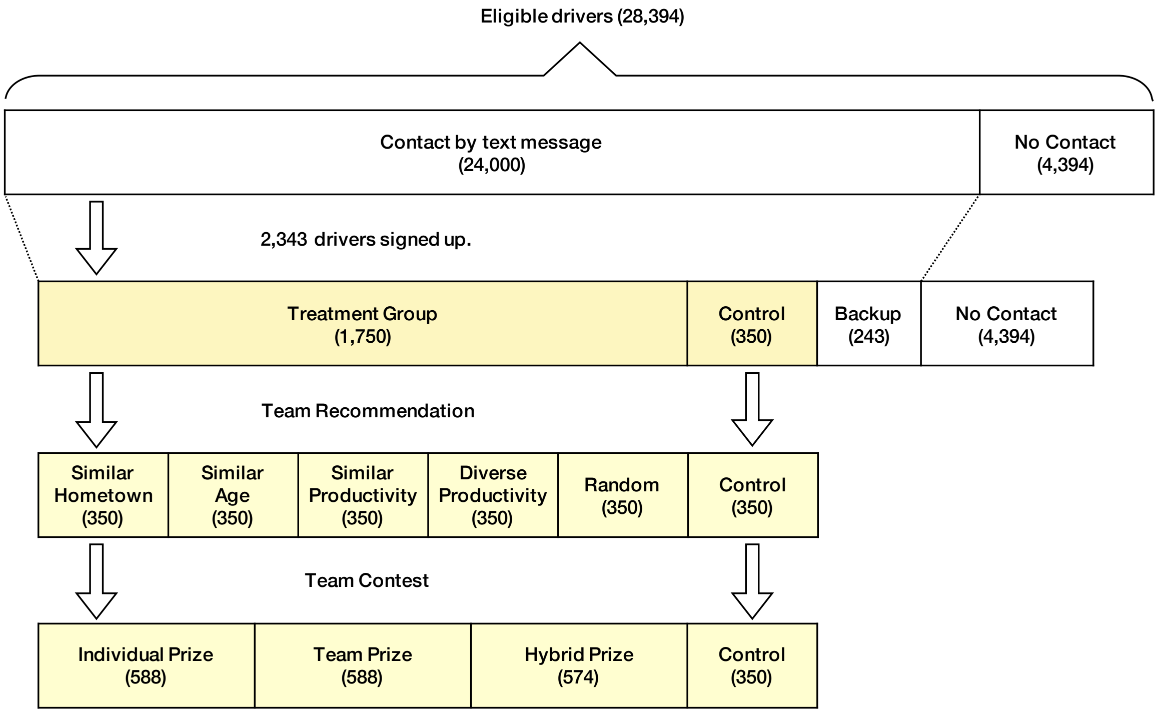


Figure 5.1: Experimental Procedure

5.3 Results

In this section, we present the results from our field experiment. We first examine the effect of our contest on overall driver productivity. We then examine our results related to the impact of team formation on team communication and performance. Finally, we end with a discussion of the effect of leadership experience through our results regarding team captain assignments.

We first investigate the average treatment effect, i.e., the effects of team contest on driver productivity. Figure 5.2 presents our results for driver productivity before, during, and after the contest period by experimental condition. The top panel presents the comparison across three experimental conditions: drivers who were never contacted (*no contact*, light dashed line), those who expressed interest but were not assigned to a team (*control*, black dashed line), and those who were assigned to a team (*treatment*, solid green line). The bottom panel further breaks the treated drivers into

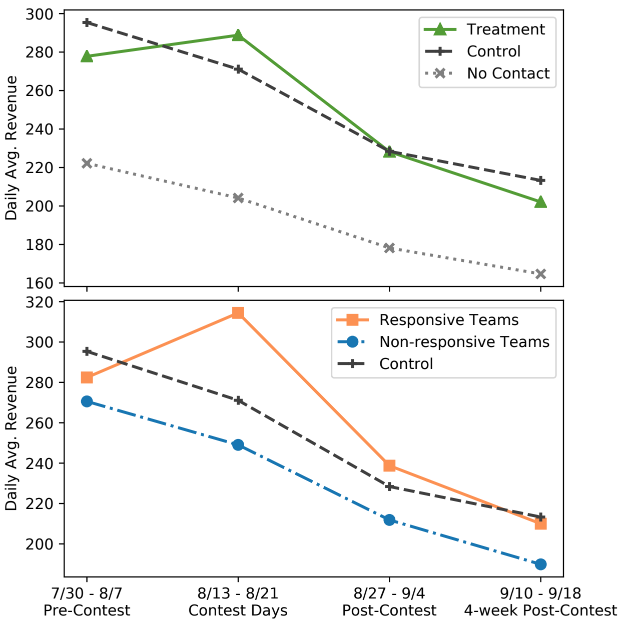


Figure 5.2: Driver Productivity Before, During, and After the Contest.

Driver productivity is measured in average daily revenue. Contest Days refer to August 13, 15, 17, 19, and 21, the dates on which the contests were conducted. We shift the dates by -14, +14, and +28 days to obtain the Pre-Contest, Post-Contest, and 4-week Post-Contest periods. Note that driver productivity is calculated only on the 5 days in each period accordingly. In the upper panel, drivers in the No Contact group are those who meet our criteria but are not randomly selected to receive an invitation to participate in our experiment. Drivers in the Treatment group are those who sign up for the experiment and are assigned to a team. Drivers in the Control group are those who sign up for the experiment but are not assigned to a team or participate in the contest. In the lower panel, we break drivers in the Treatment group into Responsive versus Non-responsive teams based on whether the team captain submits the survey.

those in responsive (solid orange line) versus non-responsive (blue dashed line) teams.

We refer to the five days of our inter-team contest as *contest days* and the 14 days prior to (post) the contest as the *pre-* (*post-*) *contest* periods. Finally, to investigate whether our effects last more than two weeks, we create a the *4-week post-contest* period. Our choice of windows ensures that we always compare the same day of the week pre-contest, contest and post-contest. During our experiment, we record daily data on each driver including the number of completed trips, the number of hours worked, and the revenue generated. On the DiDi platform, drivers receive 80% of the revenue they generate and give the remaining 20% to the platform.

Returning to Figure 5.2, we see from the upper panel that those who sign up to join a team, regardless of whether they are assigned to a treatment or control condition, are more productive than those who are never contacted (grey dashed line). Figure 5.2 also shows that both the control group and the no-contact group exhibit a similar decreasing trend over the eight-week time period of our experiment, a pattern similar to the platform's typical attrition rate. Our results in the bottom panel of Figure 5.2 show that drivers assigned to a responsive versus a non-responsive team demonstrate a large increase in revenue during the the contest period but a smaller increase in the two-week post-game period.

To quantify the average and heterogeneous treatment effects on daily revenue, we construct the following difference-in-differences models:

$$\Delta \text{Revenue}_{i,t} = \beta_0 + \beta_1 * \text{Treated} + \epsilon_{i,t}, \tag{5.1}$$

$$\Delta \text{Revenue}_{i,t} = \beta_0 + \beta_1 * \text{Responsive} + \beta_1 * \text{Unresponsive} + \epsilon_{i,t}, \tag{5.2}$$

where $\Delta\text{Revenue}_{i,t}$ represents the revenue increase of the t-th day in the current period compared to the t-th day in the pre-contest period. We report the results of these models in Table 5.2, including both the average (specifications 1-3, eq.5.1) and heterogeneous (4-6, eq.5.2) treatment effects. Pooling drivers across all treatment and control conditions, we find that the daily revenue increases by 35 CNY during the contest period compared to the pre-contest period and that this effect persists during the two-week post-contest period, albeit with half of the effect size.

Separating the results by team responsivity (specifications 4-6), we find that the increased revenue for those in a responsive team doubles the average treatment effect, whereas the treatment effect for unresponsive teams is not significantly different from

Table 5.2: Average and Heterogeneous Treatment Effects on Daily Revenue. Difference-in-differences linear regressions. We compare each of the three time periods with the Pre-Contest period.

	Dependent variable: Δ of Daily Revenue (CNY)					
	Average Treatment Effects			Heterogeneous Treatment Effects		
	(1)	(2)	(3)	(4)	(5)	(6)
Time Period	Contest	2-weeks Post Contest	4-weeks Post Contest	Contest	2-weeks Post Contest	4-weeks Post Contest
Treated	35.24***	17.36*	6.369			
	(9.319)	(9.679)	(10.09)			
	[0.001]	[0.166]	[0.68]			
Responsive				56.21***	23.25**	9.607
				(9.999)	(10.12)	(10.55)
				[0.001]	[0.066]	[0.654]
Unresponsive				2.706	8.237	1.348
				(10.14)	(10.88)	(11.23)
				[0.889]	[0.675]	[0.905]
Constant	-24.24***	-66.96***	-82.06***	-24.24***	-66.96***	-82.06***
	(7.892)	(8.844)	(9.192)	(7.892)	(8.844)	(9.193)
# Driver	2,100	2,100	2,100	2,100	2,100	2,100
Observations (Driver * Day)	10,500	10,500	10,500	10,500	10,500	10,500

Standard errors in parentheses are clustered at the contest (individual) level for treatment (control) conditions.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$
False discovery rate adjusted q-values are in square brackets.

zero.

We also examine our results controlling for demographics (Table 5.3) and correcting for multiple hypothesis testing (false discover rate adjusted q-values reported [5]) and find that our results persist. Finally, when we use the number of daily trips (Table 5.4) or the number of hours worked (Table 5.5) as our dependent variable, we find that our results again remain the same.

In our experiment, we are also interested in whether different ways of forming teams have different effects on our results. We begin by examining the effect of team formation on captain responsiveness. From Figure 5.3, we see that 39.2% of our assigned captains do not submit their questionnaires during the study period. We label these teams as our *non-responsive* group. To identify team cohesiveness (or cooperativeness), we check the accuracy of the license plate information submitted on the survey. As the DiDi platform does not contain any team communication tools, we expect that most teams communicate by phone or WeChat,[1] an expectation which is verified by our post-experiment interviews.

[1] WeChat is the dominant communication app in China, which allows group communication.

Table 5.3: Average and Heterogeneous Treatment Effects on Daily Revenue. Difference-in-differences linear panel regressions. We compare each of the three time periods with the Pre-Contest period.

	Dependent variable: Δ of Daily Revenue (CNY)					
	Average Treatment Effects			Heterogeneous Treatment Effects		
	(1)	(2)	(3)	(4)	(5)	(6)
Time Period	Contest	2-weeks Post Contest	4-weeks Post Contest	Contest	2-weeks Post Contest	4-weeks Post Contest
Treated	35.39^{***}	17.62^{*}	6.308			
	(9.306)	(9.637)	(10.04)			
	[0.001]	[0.154]	[0.682]			
Responsive				56.30^{***}	23.37^{**}	9.518
				(9.971)	(10.07)	(10.49)
				[0.001]	[0.063]	[0.634]
Unresponsive				2.900	8.700	1.320
				(10.10)	(10.83)	(11.15)
				[0.871]	[0.634]	[0.906]
Age	0.741	0.801	0.829^{*}	0.636	0.772	0.813^{*}
	(0.518)	(0.529)	(0.465)	(0.505)	(0.521)	(0.465)
DiDi Age (yr.)	17.16^{***}	18.19^{***}	6.455	17.30^{***}	18.23^{***}	6.476
	(6.606)	(6.709)	(6.252)	(6.539)	(6.711)	(6.243)
Local	14.23^{*}	3.998	24.99^{***}	15.09^{**}	4.234	25.12^{***}
	(7.621)	(7.833)	(7.657)	(7.418)	(7.838)	(7.680)
Male	35.42	27.18	34.82	34.92	27.04	34.74
	(29.03)	(29.82)	(26.08)	(29.44)	(30.10)	(26.09)
Constant	-103.4^{***}	-138.6^{***}	-157.0^{***}	-99.57^{***}	-137.5^{***}	-156.4^{***}
	(36.56)	(36.71)	(30.96)	(36.71)	(36.82)	(31.09)
# Driver	2,100	2,100	2,100	2,100	2,100	2,100
Observations (Driver * Day)	10,500	10,500	10,500	10,500	10,500	10,500
H_0: Responsive = Unresponsive				$p < 0.001$	$p = 0.0673$	

Standard errors in parentheses are clustered at the contest (individual) level for treatment (control) conditions.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$
False discovery rate adjusted q-values are in square brackets.

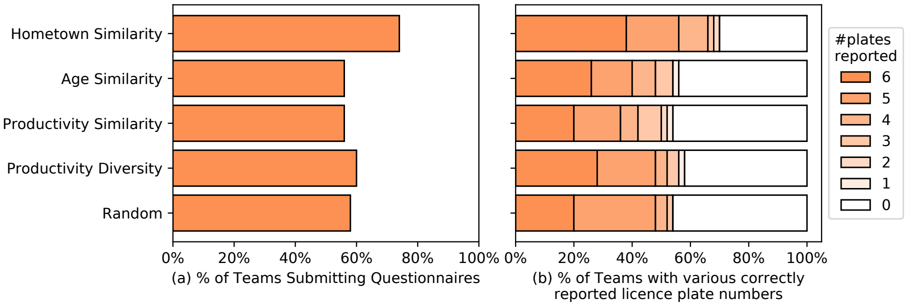


Figure 5.3: Team Responsiveness in Different Treatments.

Team Responsiveness is coded based on the pre-contest survey. Panel (a) codes the responsiveness binarily, with a team deemed responsive if the captain submits the questionnaire on team member characteristics. Panel (b) codes responsiveness based on the number of correctly-reported license plate numbers.

Table 5.4: Average and Heterogeneous Treatment Effects on Daily Trips. Difference-in-differences linear panel regressions. We compare each of the three time periods with the Pre-Contest period.

	Dependent variable: Δ of Daily Trips					
	Average Treatment Effects			Heterogeneous Treatment Effects		
	(1)	(2)	(3)	(4)	(5)	(6)
Time Period	Contest	2-weeks Post Contest	4-weeks Post Contest	Contest	2-weeks Post Contest	4-weeks Post Contest
Treated	2.392***	1.219**	0.462			
	(0.513)	(0.542)	(0.559)			
	[0.001]	[0.057]	[0.461]			
Responsive				3.493***	1.494***	0.574
				(0.555)	(0.567)	(0.584)
				[0.001]	[0.026]	[0.419]
Unresponsive				0.684	0.791	0.289
				(0.560)	(0.617)	(0.627)
				[0.334]	[0.334]	[0.646]
Constant	-2.032***	-4.408***	-5.082***	-2.032***	-4.408***	-5.082***
	(0.434)	(0.493)	(0.513)	(0.434)	(0.493)	(0.513)
# Driver	2,100	2,100	2,100	2,100	2,100	2,100
Observations (Driver * Day)	10,500	10,500	10,500	10,500	10,500	10,500
H_0: Responsive = Unresponsive				$p < 0.001$	$p = 0.1343$	

Standard errors in parentheses are clustered at the contest (individual) level for treatment (control) conditions.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$
False discovery rate adjusted q-values are in square brackets.

Table 5.5: Average and Heterogeneous Treatment Effects on Working Hours. Difference-in-differences linear panel regressions. We compare each of the three time periods with the Pre-Contest period.

	Dependent variable: Δ of Daily Working Hours					
	Average Treatment Effects			Heterogeneous Treatment Effects		
	(1)	(2)	(3)	(4)	(5)	(6)
Time Period	Contest	2-weeks Post Contest	4-weeks Post Contest	Contest	2-weeks Post Contest	4-weeks Post Contest
Treated	0.772***	0.379*	0.134			
	(0.192)	(0.197)	(0.221)			
	[0.001]	[0.125]	[0.7]			
Responsive				1.205***	0.484**	0.188
				(0.207)	(0.205)	(0.230)
				[0.001]	[0.057]	[0.623]
Unresponsive				0.0996	0.217	0.0509
				(0.207)	(0.226)	(0.248)
				[0.71]	[0.606]	[0.838]
Constant	-0.521***	-1.579***	-1.225***	-0.521***	-1.579***	-1.225***
	(0.162)	(0.180)	(0.203)	(0.162)	(0.180)	(0.203)
# Driver	2,100	2,100	2,100	2,100	2,100	2,100
Observations (Driver * Day)	10,500	10,500	10,500	10,500	10,500	10,500
H_0: Responsive = Unresponsive				$p < 0.001$	$p = 0.1148$	

Standard errors in parentheses are clustered at the contest (individual) level for treatment (control) conditions.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$
False discovery rate adjusted q-values are in square brackets.

Table 5.6: Treatment Effects on Team Responsiveness. The extensive margin measures whether the team captain submits the questionnaire, reporting the average marginal effects of Probit estimates. The intensive margin measures the number of license plates reported correctly. The omitted group is Productivity Similarity.

	Extensive margin		Intensive margin	
	Probit, $Y = P(\text{Response})$		OLS, $Y = \#\text{Correct Plates} — \text{Response}$	
	(1)	(2)	(3)	(4)
Age Similarity	0	-0.00323	0.429	0.327
	(0.0952)	(0.0964)	(0.404)	(0.412)
Hometown Similarity	0.186*	0.191**	0.437	0.403
	(0.0967)	(0.0967)	(0.378)	(0.382)
Productivity Diversity	0.0387	0.0304	0.431	0.358
	(0.0954)	(0.0956)	(0.397)	(0.402)
Random	0.0193	0.00728	0.326	0.265
	(0.0953)	(0.0957)	(0.400)	(0.405)
Avg. Daily Revenue (100 CNY)		0.0337		0.0290
		(0.0260)		(0.105)
Age		0.00303		-0.0106
		(0.00427)		(0.0175)
DiDi Age (Yr.)		-0.0119		-0.0240
		(0.0523)		(0.209)
Local		-0.0176		-0.469
		(0.0750)		(0.308)
Male		-0.000324		-0.785
		(0.225)		(0.893)
Constant			4.536***	5.769***
			(0.285)	(1.156)
Observations (# teams)	250	250	152	152
H_0: Hometown Similarity = Age Similarity	p=0.0542	p=0.0455		
	[0.079]	[0.079]		
H_0: Hometown Similarity = Random	p=0.0862	p=0.0586		
	[0.087]	[0.079]		

Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$
False discovery rate adjusted q-values are in square brackets.

Examining the results in Figure 5.3, we find that those teams based on hometown similarity are more responsive than other team formations. This result is consistent with prior research that shows that location similarity is a strong predictor of whether a member of an online community joins a team [1]. In Table 5.6, we present the results of our regression analyses. Specifications (1) and (2) use a Probit regression to examine the treatment effect along the extensive margin, with the likelihood of submitting the survey as the dependent variable. By contrast, specifications (3) and (4) use an OLS regression to examine the treatment effect along the intensive margin, with the number of license plates reported correctly as the dependent variable. The results in Table 2 again show that teams based on hometown similarity show the highest level of responsiveness. Quantitatively, these teams are 19% more likely to be responsive than age-similar teams, productivity-similar teams, or randomly-composed teams (significant at the 0.1 level), an effect that persists after controlling for demographics. Along the intensive margin, however, we do not find any significant differences among the teams which submitted the survey. One possible reason for this finding may be that the captains decide to submit their surveys only if they have sufficient information. Indeed, most captains who submit the survey get 6 or 5 plate numbers (43.4%, 31.6%) correct.

$$Pr(\text{Responsiveness}_i) = \Phi(B \cdot \text{Treatment}_i + \Gamma \cdot \text{Demographics}_i + \epsilon_i) \tag{5.3}$$

$$\#\text{Correct-Plates}_i = B \cdot \text{Treatment}_i + \Gamma \cdot \text{Demographics}_i + \epsilon_i \tag{5.4}$$

In addition to our findings on team formation and responsiveness, we are also interested in whether the type of team has an effect on driver productivity. Table 5.7 presents our results using team formation strategy as the independent variables in specifications 1-3 and team diversity as the independent variables in specifications

Table 5.7: Similarity and Diversity on Driver Productivity. DID regressions on drivers who belong to a team. Dependent variable: Difference in driver productivity (compared with the pre-contest time window). For (1-3), the omitted category is the Random treatment.

	Dependent variable: Δ Daily Revenue (CNY)					
	By Treatment Group			By Diversity Metrics		
	(1)	(2)	(3)	(4)	(5)	(6)
Time Period	Contest	2-weeks Post Contest	4-weeks Post Contest	Contest	2-weeks Post Contest	4-weeks Post Contest
Age Similarity	0.933	33.19**	9.806			
	(16.91)	(12.70)	(11.05)			
Hometown Similarity	5.838	20.70	17.12			
	(18.35)	(13.16)	(13.62)			
Productivity Similarity	-14.65	21.47*	13.85			
	(17.15)	(12.04)	(12.67)			
Productivity Diversity	-17.50	17.50	11.33			
	(15.62)	(12.25)	(13.09)			
Age Std.				-0.417	-3.357**	-0.123
				(1.647)	(1.346)	(1.279)
Avg. Hometown Distance				0.0297	-0.00706	-0.0196
				(0.0242)	(0.0227)	(0.0203)
Productivity Std.				0.0953	-0.0347	-0.00401
				(0.122)	(0.0882)	(0.0961)
DiDi Age Std.				-0.0646	-0.0370	-0.0852
				(0.0914)	(0.0852)	(0.0799)
Constant	16.07	-68.17***	-86.12***	4.701	-15.89	-48.15**
	(13.69)	(9.377)	(8.566)	(29.68)	(21.04)	(22.52)
# Driver	1,750	1,750	1,750	1,750	1,750	1,750
Observations	8,750	8,750	8,750	8,750	8,750	8,750

Standard errors in parentheses are clustered at the contest (individual) level for treatment (control) conditions.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4-6. More specifically, we measure driver diversity based on driver age, productivity, and DiDi age with their standard deviation within a team; we measure hometown diversity using the average distance (km) between any two drivers within the same team. Our results in Table 5.7 show that, irrespective of our independent variables, team formation has no significant effect on driver productivity either during (Contest) or long after (4-week Post-Contest) the contest. Interestingly, though, we find that teams based on age similarity exhibit significantly higher revenue immediately after (2-week Post-Contest) the contest, earning 33 CNY on average compared with drivers in randomly-formed teams (specification 2). This observation is confirmed by the negative correlation between age standard deviation and team productivity (specification 5).

Finally, we are interested in the productivity of those who volunteer to be captains in our study. Our results in Table 5.8 (in SI) show that those who had been more

Table 5.8: Team Captain Volunteers. Probit estimates. Reported results are average marginal effects. We include all drivers who signed up for the competition.

	Dependent Variable Volunteering to be Captain
Avg. Daily Revenue (100 CNY)	0.0253***
	(0.00760)
Male	0.0377
	(0.0600)
Local	-0.0298
	(0.0201)
Age	-0.00124
	(0.00117)
DiDi Age (years)	0.0297**
	(0.0150)
# Drivers	2,343

Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5.9: Effect of Being Appointed as a Captain: Difference-in-differences linear regressions. Dependent variable: Difference of driver productivity (compared with the pre-contest time window). Subjects are drivers who volunteer to be captains and are assigned to teams which have multiple volunteers. Note that only one volunteer in each team is randomly selected to be the captain. Kolmogorov–Smirnov tests find no significant difference in prior productivity, age, DiDi age, or gender between the the selected captains and other volunteers ($p > 0.1$).

	Dependent Variable: Δ of Daily Revenue (CNY)		
	(1)	(2)	(3)
Time Period:	Contest	2-weeks Post Contest	4-weeks Post Contest
Assigned Captain	34.181*	23.647	-5.278
	(19.534)	(19.673)	(18.624)
Constant	-17.910*	-57.146***	-65.077***
	(12.589)	(13.759)	(14.707)
# Volunteers	298	298	298
Obs.	1,490	1,490	1,490

Standard errors in parentheses are clustered at the contest (individual) level for treatment (control) conditions.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

productive in the month prior to our experiment, as well as those who had been on DiDi for a longer period of time (i.e., their DiDi age), were significantly more likely to volunteer to lead. In 141 teams with two or more drivers who express interests in being a captain, only one in each team is randomly appointed as the team captain. OLS regressions show that among our base of 298 volunteers, those who are randomly chosen to be captains are more active than those who are randomized out, earning 34 CNY more per day on average ($p < 0.1$) during contest days (Table 5.9 in SI), although this result is marginally significant.

5.4 Concluding Remarks

Our study uses a field experiment at a ride-sharing platform in China to understand how team formation and other factors impact team responsiveness and driver productivity. Applying social identity theory to the ride-sharing context, we use different team formation strategies to place drivers in teams and compare our treatment and control groups on their revenue earned during a contest. Our results show that, compared to those in the control condition, treated drivers work longer hours, complete more trips, and earn higher revenue during the contest, with a much larger effect for responsive teams who communicated more with each other prior to the inter-team contest. Overall, we find that our treated drivers earn a 12.5% higher revenue than those in the control group. Furthermore, we find that drivers in responsive teams as well as those in teams comprised of drivers with similar age continue to be more productive during the two weeks after the contest, absent of any cash prize or formal team structure.

We conclude with a few observations on how our experiment was perceived by our subjects. Our post-contest survey (see SI) and interviews indicate that over 88% of the drivers like or extremely like the team contest, citing team belonging (66%), making friends (70%), a sense of honor from winning (61%), and monetary incentives

(68%) as the top benefits. Encouraged by the results of our experiment, DiDi shipped our team-formation algorithms into production within their platform. In 2018 alone, DiDi conducted 1,548 inter-team contests across 180 cities in China, involving over two million drivers. These contests, typically one-week long, helped the platform to meet high demand from tourists during national holidays, and increased both driver income as well as DiDi's profits. While our experiment examines the effect of team formation on one platform, our results indicate that team identity shows great promise as a design tool that can be leveraged to increase worker productivity in the modern gig economy. Future research could use our study as a foundation for exploring the full potential of social identity theory, examing the impact of longer contests and more persistent teams.

5.5 Extended Materials

5.5.1 Experimental Design

We select our pool of drivers based on their productivity in a two-week period (July 18th, 2017 to July 31th, 2017). In addition to other requirements (e.g. the driver is not affiliated with a rental company), we use the following criteria to filter the drivers:

- The driver is registered in Dongguan.
- The driver has worked (finished one or more trips) on at least 5 weekdays and 2 weekend days during the two-week period.
- The driver finishes 5 or more trips on average on the days she works during the two-week period.

This filtering process yields a total of 28,394 eligible drivers. From this pool, we randomly selected 24,000 drivers to receive a treatment assignment invitation. The remaining 4,398 drivers comprise our no-contact control group. In each treatment invitation, we ask if the driver would like to sign up for "a team competition" that earns up to 1,000 CNY for the team. Additionally, we ask if a driver is interested in being a team captain, with an additional 100 CNY received upon fulfilling the responsibilities.

Our invitations received 2,343 positive responses, 531 of which were interested in being a team captain. For our experiment, we divide our positive responses into 3 sets.

- Set 1 includes 1,750 drivers and constitutes our treatment group. These drivers are subsequently partitioned into teams of 7 (250 teams in total).
- Set 2 includes 350 drivers, constituting the placebo group. These drivers are not placed in a team.
- Set 3 includes the remaining 243 drivers, constituting the "backup group." If a driver in the treatment group drops out before the competition, we replace the drop-out driver with a similar driver in the backup group. Indeed, in our experiment, 15 drivers were reported by their captain as not responsive or no longer available for competition. We mark these 15 drivers as "dropouts" and replace them with similar drivers from the backup group.

During the team formation stage, we first randomly partition the 1,750 drivers from Set 1 into five conditions. We then group drivers in each condition into 50 teams with the same strategy. For example, in the *Similar-Hometown* condition, the seven drivers in the same team are all from the same (or a nearby) province. For each team, we ensured that at least one member has volunteered to be a team captain.

After forming the teams, we next text each team captain the phone numbers of the drivers in their team. We also text each team member the phone number of their team captain. In addition, we ask team captains to fill out a questionnaire designed to verify team formation and provide possibilities for icebreaking communications within the team. The questionnaire includes the following questions:

- What are the last 3 digits of the plate numbers of the team members (six blanks, excluding the team captain's)?
- What is the name of your team (after discussing with the team members)?
- Where is the farthest hometown (from Dongguan) of your team members?
- What is the maximum age of the team members?

In the team competition stage, each team is paired with another team of similar productivity throughout the duration of the experiment period. We use a 3×2 design to vary the prize structure and in-team coordination. For the prize structure, drivers earn monetary awards based on their individual performance, team performance, or both. Along the coordination axis, we either use a system to determine their position for the next game day, or we allow teams to adjust their own positions. Since they do not affect our major outcome, we eliminate them in the analysis.

5.5.2 Power Analysis

To determine the number of drivers needed in our experiment, we conduct a pilot experiment in a different city and find that drivers who are willing to participate in a team contest complete 11.7 orders on average per day (std. 4.7). Since we expect an effect size of 10%, with $\alpha = 0.05$ and 0.9 power, this requires us to have 170 drivers per treatment. If we assume that 50% drivers who sign up for the experiment will complete the experiment, we need 340 drivers per treatment. This leads us to selecting a subject pool of 350 drivers per treatment.

5.5.3 Survey Response

The response rate is 577/1750=33%. The number and percent of drivers choosing a certain choice are indicated in the bracket.

1. Did you participate in the team contest in XXX city from XX to XX?

   (a) Yes. (99.%)

   (b) I am not sure. (1%)

2. To what degree did you enjoy this team contest? Please rate on the scale below from 1 (extremely dislike) to 5 (extremely like).

   (1) 8 (1.4%)

   (2) 8 (1.4%)

   (3) 53 (9.2%)

   (4) 80 (13.9%)

   (5) 428 (74.2%)

3. Why do you like this team contest? Please select all that apply. (Limited to the 508 drivers who chose 4 or 5 in Q2.

   (a) I had a sense of team belonging. (337, 66.3%)

   (b) The contest was interesting and thrilling. (241, 47.4%)

   (c) I was able to make more friends. (334, 65.7%)

   (d) Winning the contest gave me a sense of honor. (310, 61.0%)

   (e) I got a monetary bonus. (347, 68.3%)

   (f) Other reasons, please specify: (20, 3.9%)

4. Why did you dislike the team contest? Please select all that apply. (Limited to the 69 drivers who chose 1, 2, or 3 in Q2.)

(a) The team members were not collaborative or united enough. (30, 43.5%)

(b) The team was not active enough to justify existence. (42, 60.9%)

(c) The leader didn't have good leadership and management skills. (31, 44.9%)

(d) The contest rules were too complicated for me to understand. (4, 5.8%)

(e) The contest rules were unfair. (10, 14.5%)

(f) The monetary bonus was not large enough to attract me. (38, 55.1%)

(g) Other reasons, please specify: (9, 13.0%)

5. As a team [member/captain] , how did you benefit from this team contest? Select all that apply.

(a) I made more friends. (405, 70.2%)

(b) I improved my leadership skills. (only for captains, 82, 68.9% among captains)

(c) I improved my communication skills. (278, 48.2%)

(d) I improved my collaboration skills with other drivers. (342, 59.3%)

(e) I became more experienced and skillful about taking the DiDi orders. (300, 52.0%)

(f) I received consolation from my teammates when I was down. (191, 33.1%)

(g) Other reasons, please specify. (33, 5.7%)

6. Which of the rules in this contest do you like? Please select all that apply.

(a) There was one day off between every two contest days. (270, 46.8%)

(b) The score was announced immediately after each contest day. (315, 54.6%)

(c) There were both driver-level and team-level competition. (402, 69.7%)

(d) The team could discuss and decide the lineup together. (66, 31.6% among the 209 applicable participants)

(e) The lineup changed between contest days. (195, 33.8%)

(f) Other reasons, please specify: (7, 1.2%)

(g) None (22, 3.8%)

7. How did your team do in this contest? Please select all that apply

   (a) Although each team member was different, we all got along well. (307, 53.2%)

   (b) Our team shared commonalities and common topics. (268, 46.4%)

   (c) Everyone contributed to our team's honor during the contest. (398, 69.0%)

   (d) Inactive team members influenced others' enthusiasm for the contest. (186, 32.2%)

8. What would you choose if you participate the contest again?

   (a) I would choose to be a team member. (359, 62.2%)

   (b) I would choose to be a team captain. (158, 27.4%)

   (c) I haven't decided. (60, 10.4%)

9. Why did you prefer NOT to be a team captain? (Applicable only to drivers who chose to be team member in Q8.)

   (a) I didn't want to initiate communication with strangers. (12, 5.5%)

   (b) I didn't know how to lead a team. (90, 41.3%)

   (c) The extra bonus for team captains was not enough. (28, 12.8%)

   (d) Team captains required too much extra work. (54, 24.8%)

(e) I am inexperienced with team management and I need more practice. (130, 59.6%)

(f) Other reasons, please specify: (12, 5.5%)

10. What do you think a team captain should do?

    (a) Lead by example. (409, 70.9%)

    (b) Be positive and energetic. (379, 65.7%)

    (c) Help teammates be more active. (416, 72.1%)

    (d) Help the team to win the contest. (372, 64.5%)

    (e) Represent team members with feedback and suggestions to the DiDi platform. (329, 57.0%)

    (f) Other reasons, please specify: (12, 2.1%)

11. How did you prefer to join team?

    (a) I preferred to join the WeChat group of my team and communicate with other teammates online. (71, 12.3%)

    (b) I preferred to call others and ask to join their team. (316, 54.8%)

    (c) I preferred to wait for a phone-call invitation to join the team. (186, 32.2%)

    (d) Other reasons, please specify: (4, 0.7%)

12. Which of the following teams would you prefer?

    (a) Temporary teams, so that I can join different teams for each contest. (116, 20.1%)

    (b) A long-lasting team, and team members keep in touch after the contest. (461, 79.9%)

13. Which of the following team structures would you prefer?

(a) I prefer to have a captain and different roles among team members. (162, 28.1%)

(b) I don't care if there is a team captain, as long as all teammates can work hard together. (412, 71.4%)

(c) Others, please specify: (3, 0.5%)

14. Do you have any other suggestions for team activities?

CHAPTER VI

Conclusion

This dissertation is broadly aligned with data science for social good, which applies data science to address real-world challenges for societal benefit. In many cases, social good can be achieved simply through a bottom-up effort if individuals adopt pro-social behaviors.

However, persuading users to adopt pro-social behaviors is hard, especially when there isn't a clear or immediate incentive to do so. Instead of enforcing policies, we seek to use so-called *nudge* behavior mechanisms, which have demonstrated effectiveness in several applications. Specifically, this dissertation seeks to better nudge with an end-to-end data science pipeline.

Ideally, this pipeline includes three interrelated stages: (1) finding, (2) implementing, (3) and evaluating a nudge, all conducted in a data-driven approach, especially with the joint forces of causal inference and machine learning. Below, we layout the ideal setup of the three stages, and map them to the three application scenarios described in Chapters III to V. Note that some of the works mentioned below may not have been done for this dissertation. Either they have been done in previous studies, or they are not immediately feasible and thus planned as future work. Yet this does not render the projects incomplete but instead highlights the flexibility of our proposed pipeline.

- Causal inference, with the help from machine learning, can be used on user-behavioral data, collected either from data records or from randomized experiments, to reveal causal insights about potential nudges for pro-social behavior: On GitHub, we identify using emoji as the nudge to more participation and effort in resolving issues; On Kiva and DiDi, we identify joining teams as the nudge to higher contributions or productivity.

- Recommender systems provide personalized suggestions for each individual. By providing the users with more relevant choices, the recommender systems maximize the effect of the identified nudge and increase the adoption of pro-social behaviors. On GitHub, such a recommender system would recommend the emoji that the users are most likely to use in submitting an issue. On Kiva and DiDi, the recommender system would recommend the teams that the drivers are more likely to join.

- Finally, randomized field experiments put the recommender systems into practice and evaluate them in real-world contexts. On GitHub, this would test the emoji recommender system as to how the recommendations are accepted and whether the increased usage of emoji (if any), triggers more attention and participation. On Kiva and DiDi, they would see if users are more likely to join teams and become more active.

It should also be noted that output from the last stage could also serve as input to the first stage, initializing a new iteration of the pipeline.

6.1 Implications

The proposed end-to-end pipeline presents implications to audiences in several different disciplinaries:

- For the machine learning audience, the causal questions raised in the pipeline, that is predicting the causal effect of a treatment, substantially change the setup of supervised machine learning. The outcome variable to be predicted, the causal effect, can never be truly observed, as it requires comparison with a counterfactual outcome. One should be aware that the valid estimate of causal effect requires different assumptions associated with causal inference techniques. In most cases, there are assumptions that cannot be fully justified with data or metrics.

- For causal inference researchers, the power of machine learning to handle large and complex data sets extends the ability to make previously impossible causal conclusions. Not only can we apply off-the-shelf machine learning algorithms to extract features from images, texts, or other complex data types, but we can also adapt machine learning algorithms to existing identification strategoies and estimate both average treatment effects and heterogeneous treatment effects. Note that the latter application usually requires modifying existing algorithms to provide valid confidence intervals for the estimation. One should be aware that machine learning models are all "data-driven," and frequently use cross-validation for model selection and parameter tuning.

- For domain practitioners, our proposed pipeline provides a principled way to integrate domain knowledge with advanced data science techniques. Such domain knowledge could be developed theory in literature, conclusions drawn from previous studies, or heuristics based on daily practices. In the first stage, domain knowledge provides the initial hypothesis of the potential nudges, which can be tested on real-world data. In the second stage, we rely on domain knowledge for useful features that help improve the performance of the recommender system. Finally, in the third stage, domain knowledge is critical in designing a well-controlled field experiment.

6.2 Future Direction

Our proposed end-to-end data science pipeline is principled and flexible. Not only is it easy to apply on different platforms, it is also easy to extend existing explorations further down the pipeline. In the last part of this dissertation, we outline a few extensions for the three applications described in the early chapters, and detail a few additional applications.

Promote Developer Participation with Emoji Our exploration of GitHub started at the first stage. With propensity score matching, we successfully identified emoji as potential means to promote developer participation. The immediate next stage is to develop an emoji recommender system to encourage the use of emoji on the platform. In fact, we have already seen efforts from GitHub to ease the typing of emoji. For example, GitHub's online editor already supports transforming the `:emoji_alias:` formatted emoji aliases into emoji characters. For example, by typing `:+1:`, the interface would immediately suggest 👍 for the user to click and insert into the text.

Team Recommendation on Kiva In a sense, our study on team recommendations on Kiva has already gone through an entire pipeline, except that the first stage was completed in a previous study. However, if we consider the perspective of the teams instead of the users, the Kiva users recommended to join a team would also serve as a stimulus to the recommended team. In other words, the same set of experimental data can also be used to analyze the newcomers' effect on teams. That is, we are re-iterating the pipeline from the first stage. In fact, a pilot study has revealed significant variations in team activity levels. If we conclude that an active newcomer would actually re-activate a dormant team, that would be a nudge to promote the lending of the Kiva users who are members of only inactive teams. Further down the pipeline, we may adapt the team recommender system to recommend active newcomers to inactive teams.

Team Competition on DiDi In the team competition conducted on DiDi, we used a rule-based system to partition participants into teams, and we start at the third stage, field experiment, on our pipeline. With the experiment data, however, we are able to re-iterate the pipeline and build data-driven team recommender systems. Such systems can then be used to support more field experiments on DiDi, and further the exploration of the driver team mechanism. We envision two potential future projects with the help of the renovated team recommender system: First, we may extend the team identity to beyond the short-term competition and study how team identities can enhance drivers' experiences in the long term. Such a study is the third stage in the second run of the pipeline. Second, if the team recommender system and the team competition are delivered into the product, there would be many short competitions conducted across different regions and time periods. The results from these short competitions would enable the first stage in the third iteration of the pipeline, to explore the heterogeneity observed among the drivers and competitions. In short, we may well expect several loops through the end-to-end pipeline.

Aside from extensions to the current applications, we may also apply the pipeline to other applications. Let's use online education as an example:

End-to-end pipeline to promote student outcomes in online education Recent developments in educational technology, such as MOOCs and digitized learning content, present an unprecedented opportunity for data science and education researchers: The availability of website clickstreams, forum participation, and course content interactions creates a heterogeneous data repository to model students' learning behavior. In the absence of the traditional classroom, randomized experiments can be conducted at the student level. In short, such a platform can benefit from our pipeline in exploring effective nudges to improve student outcomes. We do note that the scenarios present new challenges to the end-to-end pipeline: First, experiments in education have a long cycle, and the effect of treatments won't be observed until the end of a course or even years after a program is completed. This challenge sets a higher requirement for the first two stages in the pipeline: finding causal insights and deriving robust prediction models. Second, students intrinsically have different learning paces and backgrounds, calling for advanced tools for analyzing heterogeneity in treatment outcomes.

Give this command the relative path to the .bib file.

BIBLIOGRAPHY

[1] Wei Ai, Roy Chen, Yan Chen, Qiaozhu Mei, and Webb Phillips. Recommending teams promotes prosocial lending in online microfinance. *Proceedings of the National Academy of Sciences*, 113(52):14944–14948, December 2016.

[2] Wei Ai, Xuan Lu, Xuanzhe Liu, Ning Wang, Gang Huang, and Qiaozhu Mei. Untangling Emoji Popularity through Semantic Embeddings. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media - ICWSM '17*, pages 2–11, 2017.

[3] George A. Akerlof and Rachel E. Kranton. Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753, 2000.

[4] George A. Akerlof and Rachel E. Kranton. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being.* Princeton University Press, 2010.

[5] Michael L. Anderson. Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103(484):1481–1495, December 2008.

[6] James Andreoni. Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy*, 97(6):1447–1458, 1989.

[7] James Andreoni. Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401):464–477, 1990.

[8] James Andreoni. Toward a theory of charitable fund-raising. *Journal of Political Economy*, 106(6):1186–1213, 1998.

[9] James Andreoni and B. Douglas Bernheim. Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects. 77(5):1607–1636, 2009.

[10] James Andreoni and A. Abigail Payne. Charitable giving. In *Handbook of Public Economics*, volume 5, pages 1–50. 2013.

[11] J. D. Angrist and A. B. Keueger. Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, November 1991.

[12] Joshua D Angrist. Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review*, 80(3):313–336, 1990.

[13] Dan Ariely, Anat Bracha, and Stephan Meier. Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *American Economic Review*, 99(1):544–555, March 2009.

[14] Susan Athey and Guido Imbens. Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, July 2016.

[15] Susan Athey and Guido W. Imbens. The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, 31(2):3–32, May 2017.

[16] Peter C. Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3):399–424, May 2011.

[17] Richard P. Bagozzi and Utpal M. Dholakia. Open Source Software User Communities: A Study of Participation in Linux User Groups. *Management Science*, 52(7):1099–1115, July 2006.

[18] Jon Bakija and Bradley T. Heim. How does charitable giving respond to incentives and income? New estimates from panel data. In *Economic Analysis of Tax Expenditures*. National Tax Journal,(National Tax Association), Vol. 64, no. 2, part 2, 2011.

[19] Eytan Bakshy, Dean Eckles, and Michael S. Bernstein. Designing and Deploying Online Field Experiments. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 283–292, Seoul, Korea, 2014.

[20] Eytan Bakshy, Dean Eckles, Rong Yan, and Itamar Rosenn. Social Influence in Social Advertising: Evidence from Field Experiments. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 146–161, Valencia, Spain, 2012.

[21] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. What does this emoji mean? A vector space skip-gram model for Twitter emojis. *Proceedings of Language Resources and Evaluation Conference*, pages 3967–3972, 2016.

[22] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50, May 2014.

[23] Theodore Bergstrom, Lawrence Blume, and Hal Varian. On the private provision of public goods. *Journal of Public Economics*, 29(1):25–49, February 1986.

[24] Marianne Bertrand, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman. What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment. *The Quarterly Journal of Economics*, 125(1):263–306, February 2010.

[25] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

[26] Gary Bornstein, Uri Gneezy, and Rosmarie Nagel. The effect of intergroup competition on group coordination: An experimental study. *Games and Economic Behavior*, 41(1):1–25, October 2002.

[27] Léon Bottou, Jonas Peters, Peters Ch, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

[28] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72, February 2008.

[29] Gary Charness and Patrick Holder. Charity in the laboratory: Matching, competition, and group identity. *Management Science*, 2018.

[30] Gary Charness, Luca Rigotti, and Aldo Rustichini. Individual Behavior and Group Membership. *American Economic Review*, 97(4):1340–1352, September 2007.

[31] Roy Chen and Yan Chen. The potential of social identity for equilibrium selection. *American Economic Review*, 101(6):2562–89, 2011.

[32] Roy Chen, Yan Chen, Yang Liu, and Qiaozhu Mei. Does team competition increase pro-social lending? Evidence from online microfinance. *Games and Economic Behavior*, 101:311–333, 2017.

[33] Yan Chen, F Maxwell Harper, Joseph Konstan, and Sherry Xin Li. Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens. *American Economic Review*, 100(4):1358–1398, September 2010.

[34] Yan Chen, Teck Hua Ho, and YONG MI Kim. Knowledge Market Design: A Field Experiment at Google Answers. *Journal of Public Economic Theory*, 12(4):641–664, July 2010.

[35] Yan Chen and Joseph Konstan. Online field experiments: A selective survey of methods. *Journal of the Economic Science Association*, 1(1):29–42, 2015.

[36] Yan Chen and Sherry Xin Li. Group Identity and Social Preferences. *The American Economic Review*, 99(1):431–457, 2009.

[37] Yan Chen, Sherry Xin Li, Tracy Xiao Liu, and Margaret Shih. Which hat to wear? Impact of natural identities on coordination and cooperation. *Games and Economic Behavior*, 84:58–86, March 2014.

[38] Victor Chernozhukov, Christian Hansen, and Martin Spindler. Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments. *American Economic Review*, 105(5):486–490, May 2015.

[39] Peter Cohen, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalfe. Using Big Data to Estimate Consumer Surplus: The Case of Uber. Working Paper 22627, National Bureau of Economic Research, September 2016.

[40] Alain Cohn, Ernst Fehr, and Michel Andre Marechal. Business culture and dishonesty in the banking industry. 516(7529):86–89, 2014.

[41] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, pages 32–41, Honolulu, Hawaii, USA, 2007.

[42] Rachel Croson, Melanie Marks, and Jessica Snyder. Groups work for women: Gender and group identity in the provision of public goods. *Negotiation Journal*, 24(4):411–427, 2008.

[43] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1277–1286, Seattle, Washington, USA, 2012.

[44] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words! Linguistic Style Accommodation in Social Media. In *Proceedings of the 20th International Conference on World Wide Web - WWW '11*, page 745, New York, New York, USA, 2011. ACM Press.

[45] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of Power: Language Effects and Power Differences in Social Interaction. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 699–708, Lyon, France, 2012.

[46] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: User Lifecycle and Linguistic Change in Online Communities. In *Proceedings of the 22nd International*

Conference on World Wide Web - WWW '13, pages 307–318, New York, New York, USA, 2013. ACM Press.

[47] Stefano DellaVigna, John A. List, and Ulrike Malmendier. Testing for Altruism and Social Pressure in Charitable Giving. *The Quarterly Journal of Economics*, 127(1):1–56, February 2012.

[48] Daantje Derks, Arjan E R Bos, and Jasper von Grumbkow. Emoticons and Online Message Interpretation. *Social Science Computer Review*, 26(3):379–388, August 2008.

[49] Daantje Derks, Arjan E. R. Bos, and Jasper von Grumbkow. Emoticons in Computer-Mediated Communication: Social Motives and Social Context. *CyberPsychology & Behavior*, 11(1):99–101, 2008.

[50] Catherine C. Eckel and Philip J. Grossman. Rebate versus matching: Does how we subsidize charitable contributions matter? *Journal of Public Economics*, 87(3-4):681–701, 2003.

[51] Catherine C. Eckel and Philip J. Grossman. Managing diversity by creating team identity. *Journal of Economic Behavior & Organization*, 58(3):371–392, November 2005.

[52] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. Emoji2vec: Learning Emoji Representations from their Description. 2016.

[53] Ido Erev, Gary Bornstein, and Rachely Galili. Constructive Intergroup Competition as a Solution to the Free Rider Problem: A Field Experiment. *Journal of Experimental Social Psychology*, 29(6):463–478, November 1993.

[54] Qiang Fu, Jingfeng Lu, and Yue Pan. Team Contests with Multiple Pairwise Battles. *American Economic Review*, 105(7):2120–2140, July 2015.

[55] Matt Garley and Julia Hockenmaier. Beefmoves: Dissemination, Diversity, and Dynamics of English Borrowings in a German Hip Hop Forum. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 135–139, July 2012.

[56] M Girvan and M E J Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002.

[57] Georgios Gousios. The GHTorrent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 233–236, San Francisco, CA, USA, 2013. IEEE Press.

[58] Jonathan V. Hall and Alan B. Krueger. An Analysis of the Labor Market for Uber's Driver-Partners in the United States. *ILR Review*, 71(3):705–732, May 2018.

[59] Glenn W Harrison and John A List. Field Experiments. *Journal of Economic Literature*, 42(4):1009–1055, November 2004.

[60] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, pages 241–250, New York, NY, USA, 2000.

[61] Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[62] Tianran Hu, Han Guo, Hao Sun, Thuy-vy Thi Nguyen, and Jiebo Luo. Spice up Your Chat: The Intentions and Sentiment Effects of Using Emoji. In *Proceedings of the 11th International AAAI Conference on Web and Social Media - ICWSM '17*, 2017.

[63] Guido W Imbens. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The review of economics and statistics*, 86(1):4–29, 2006.

[64] Guido W Imbens and Jeffrey M Wooldridge. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1):5–86, March 2009.

[65] G.W. Imbens and J.D. Angrist. Identification and estimation of local average treatment effects. 62(2):467–475, 1994.

[66] Grace Youngjoo Jeon, Yong-mi Kim, and Yan Chen. Re-examining price as a predictor of answer quality in an online q&a site. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, page 325, New York, New York, USA, 2010. ACM Press.

[67] T. Joachims and A. Swaminathan. Tutorial on counterfactual evaluation and learning for search, recommendation and ad placement. In *ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1199–1201, 2016.

[68] Fredrik Johansson, Uri Shalit, and David Sontag. Learning Representations for Counterfactual Inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.

[69] M I Jordan and T M Mitchell. Machine learning: Trends, perspectives, and prospects. 349(6245):255–260, July 2015.

[70] Dean Karlan and John A List. Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment. *American Economic Review*, 97(5):1774–1793, 2007.

[71] Phillip H. Kim and Howard E. Aldrich. Teams that Work Together, Stay Together: Resiliency of Entrepreneurial Teams. SSRN Scholarly Paper ID 1768134, Social Science Research Network, Rochester, NY, 2006.

[72] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction Policy Problems. *American Economic Review*, 105(5):491–495, May 2015.

[73] Kai Konrad. *Strategy and Dynamics in Contests*. Oxford University Press, 2009.

[74] Hema A. Krishnan, Alex Miller, and William Q. Judge. Diversification And Top Management Team Complementarity: Is Performance Improved Teams? *Strategic Management Journal*, 18(May 1997):361–374, 1997.

[75] Franklin B. Krohn. A Generational Approach to Using Emoticons as Nonverbal Communication. *Journal of Technical Writing and Communication*, 34(4):321–328, October 2004.

[76] Brian K. Lee, Justin Lessler, and Elizabeth A. Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, February 2010.

[77] David S. Lee. Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, 142(2):675–697, February 2008.

[78] John List, Sally Sadoff, and Mathis Wagner. So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14(4):439–457, March 2011.

[79] John A. List and David Lucking-Reiley. The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign. *Journal of Political Economy*, 110(1):215–233, February 2002.

[80] John A. List, Azeem M. Shaikh, and Yang Xu. Multiple hypothesis testing in experimental economics. *Experimental Economics*, January 2019.

[81] Tracy Liu, Zhixi Wan, and Chenyu Yang. The Efficiency of A Dynamic Decentralized Two-Sided Matching Market. SSRN Scholarly Paper ID 3339394, Social Science Research Network, Rochester, NY, February 2019.

[82] Xuanzhe Liu, Wei Ai, Huoran Li, Jian Tang, Gang Huang, Feng Feng, and Qiaozhu Mei. Deriving User Preferences of Mobile Apps from Their Management Activities. *ACM Transactions on Information Systems*, 35(4):1–32, July 2017.

[83] Yang Liu, Roy Chen, Yan Chen, Qiaozhu Mei, and Suzy Salib. "I loan because...": Understanding Motivations for Pro-Social Lending. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining - WSDM '12*, page 503, New York, New York, USA, 2012. ACM Press.

[84] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. Learning from the ubiquitous language: An Empirical Analysis of Emoji Usage of Smartphone Users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*, pages 770–780, New York, New York, USA, 2016. ACM Press.

[85] Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

[86] Leslie M. Marx and Steven A. Matthews. Dynamic Voluntary Contribution to a Public Project. *The Review of Economic Studies*, 67(2):327–358, April 2000.

[87] Nora McDonald and Sean Goggins. Performance and participation in open source software on GitHub. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*, page 139, New York, New York, USA, 2013. ACM Press.

[88] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[89] Jonathan Meer and Harvey S. Rosen. The ABCs of charitable solicitation. *Journal of Public Economics*, 95(5):363–371, June 2011.

[90] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[91] Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. Understanding Emoji Ambiguity in Context: The Role of Text in Emoji-Related Miscommunication. *Proceedings of the 11th International AAAI Conference on Web and Social Media - ICWSM '17*, pages 152–161, 2017.

[92] Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. "Blissfully Happy" or "Ready to Fight": Varying Interpretations of Emoji. In *Tenth International AAAI Conference on Web and Social Media*, March 2016.

[93] Kate G. Niederhoffer and James W. Pennebaker. Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology*, 21(4):337–360, December 2002.

[94] S. Onoue, H. Hata, and K. Matsumoto. A Study of the Characteristics of Developers' Activities in GitHub. In *2013 20th Asia-Pacific Software Engineering Conference (APSEC)*, volume 2, pages 7–12, December 2013.

[95] Scott E. Page. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies (New Edition).* Princeton University Press, 2007.

[96] Jan Potters, Martin Sefton, and Lise Vesterlund. After you—endogenous sequencing in voluntary contribution games. *Journal of Public Economics*, 89(8):1399–1419, August 2005.

[97] Eric S. Raymond. *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary.* O'Reilly, Beijing ; Cambridge, Mass, 1st ed edition, 1999.

[98] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to Recommender Systems Handbook. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 1–35. Springer US, Boston, MA, 2011.

[99] Jeffrey A. Roberts, Il-Horn Hann, and Sandra A. Slaughter. Understanding the Motivations, Participation, and Performance of Open Source Software Developers: A Longitudinal Study of the Apache Projects. *Management Science*, 52(7):984–999, July 2006.

[100] James M. Robins and Andrea Rotnitzky. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, 90(429):122–129, March 1995.

[101] Paul R Rosenbaum and Donald B Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. 70(1):41, April 1983.

[102] Paul R. Rosenbaum and Donald B. Rubin. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

[103] Martin Ruef, Howard E Aldrich, and Nancy M Carter. The Structure of Founding Teams: Homophily, Strong Ties, and Isolation among U.S. Entrepreneurs. *American Sociological Review*, 68(2):195, April 2003.

[104] Moses Shayo. A Model of Social Identity with an Application to Political Economy: Nation, Class, and Redistribution. *American Political Science Review*, 103(2):147–174, May 2009.

[105] Vinayak Sinha, Alina Lazar, and Bonita Sharif. Analyzing Developer Sentiment in Commit Logs. In *Proceedings of the 13th International Conference on Mining Software Repositories*, MSR '16, pages 520–523, Austin, Texas, 2016.

[106] Jeffrey Smith. *Matching and Weighting Lecture*. 2014.

[107] Jeffrey A. Smith and Petra E. Todd. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1):305–353, March 2005.

[108] Elizabeth A. Stuart, Brian K. Lee, and Finbarr P. Leacy. Prognostic score–based balance measures for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology*, 66(8 0):S84–S90.e1, August 2013.

[109] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3632–3642. Curran Associates, Inc., 2017.

[110] Henri Tajfel, John C. Turner, William G. Austin, and Stephen Worchel. An integrative theory of intergroup conflict. *Organizational identity: A reader*, pages 56–65, 1979.

[111] Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 35, pages 175–185, Stroudsburg, PA, USA, September 2014. Association for Computational Linguistics.

[112] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15*, pages 1067–1077, New York, New York, USA, 2015. ACM Press.

[113] Jason T. Tsay, Laura Dabbish, and James Herbsleb. Social media and success in open source projects. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion - CSCW '12*, page 223, New York, New York, USA, 2012. ACM Press.

[114] Lise Vesterlund. The informational value of sequential fundraising. *Journal of Public Economics*, 87(3):627–657, March 2003.

[115] Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, pages 1–15, June 2018.

[116] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. Learning to Rank with Selection Bias in Personal Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '16*, pages 115–124, 2016.

[117] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3589–3597, Sydney, NSW, Australia, 2017. JMLR.org.

[118] Daniel Westreich, Justin Lessler, and Michele Jonsson Funk. Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833, August 2010.

[119] Teng Ye, Sangseok You, and Lionel P. Robert Jr. When Does More Money Work? Examining the Role of Perceived Fairness in Pay on the Performance Quality of Crowdworkers. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[120] Y. Connie Yuan and Geri Gay. Homophily of Network Ties and Bonding and Bridging Social Capital in Computer-Mediated Distributed Teams. *Journal of Computer-Mediated Communication*, 11(4):1062–1084, 2006.

[121] Rui Zhou, Jasmine Hentschel, and Neha Kumar. Goodbye Text, Hello Emoji. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, pages 748–759, 2017.