Three Essays on Digital Economics

Shaojia Wang

A Thesis
In the Department
Of
Economics

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Economics)
at Concordia University
Montréal, Québec, Canada

December 2025

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: **Shaojia Wang**

Entitled: **Three Essays on Digital Economics**

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (ECONOMICS)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________ Chair
Panos Margaris

______________________________ External Examiner
Junnan He

______________________________ Examiner
Jorgen Hansen

______________________________ Examiner
Julius Owusu

______________________________ Examiner
Gabrielle Vasey

______________________________ Thesis Supervisor
Jan Victor Dee

Approved by ______________________________
Christian Sigouin

[Month/day/year] ______________________________
Pascale Sicotte
Faculty of Arts and Science

Abstract

Three Essays on Digital Economics

Shaojia Wang, Ph.D.
Concordia University, 2025

This dissertation comprises three chapters on digital economics, examining how digital platforms—both public and private—reshape markets, consumer behaviour, and economic welfare in the digital era. The first chapter investigates the role of a government-initiated e-commerce platform (GEP) in transforming the sales of a traditional agricultural specialty in China's Pu'er tea market. Drawing on a unique dataset from field experiments and surveys of 983 farmers, it identifies significant substitution effects from offline to online sales, highlighting how the GEP's bundled public services, such as cooperative packaging, regional branding, and logistics coordination, facilitate farmers' digital participation and market expansion. The second chapter explores the renewal behaviour of digital content memberships on a Chinese creator platform, employing a reduced-form empirical analysis to quantify the impacts of price adjustments and peer influence within complex referral networks. The findings reveal that referee-targeted discounts are the most effective strategy for promoting renewals while minimizing revenue loss. Building on this evidence, the third chapter develops a structural model that endogenizes the formation of referral networks and captures users' interdependent renewal decisions, enabling counterfactual simulations of alternative pricing and referral strategies. Collectively, these chapters advance our understanding of how digitalization transforms both producer and consumer behaviour, offering theoretical and empirical insights for platform governance, public digital initiatives, and policy interventions that aim to foster inclusive, efficient, and sustainable digital economic development.

Contribution of Authors

This dissertation is a manuscript-based thesis composed of three chapters, all of which are based on co-authored research.

Chapter 1, "Digital Revitalization or Useless Effort? Public E-commerce Support and Local Specialty Sales," is co-authored with Xintong Han, Jan Victor Dee, and Kefan Chen. A version of this chapter has already been published in the Journal of Development Economics. My contribution to this project included helping develop the research idea, coordinating and conducting the field data collection with local government support, curating the household-level dataset, and carrying out the formal empirical analysis. In particular, I participated in organizing survey teams and conducting door-to-door interviews with nearly 1,000 households in Pu'er City over a two-year period, which formed the basis of the dataset used in this chapter. I also participated in robustness checks and heterogeneity analyses and contributed to revising the manuscript.

Chapters 2 and 3 are adapted from the co-authored research project with Xintong Han, Xin (Shane) Wang, and Tong Wang on referral networks and digital membership renewal. In the dissertation, this joint work is presented in two parts: Chapter 2 presents the reduced-form empirical analysis, while Chapter 3 presents the structural model and counterfactual analysis. My contribution to this project included working with the proprietary platform data, contributing to the empirical and structural analyses, interpreting the findings, and contributing to the writing and revision of the manuscript.

The dissertation introduction, the integration of the three chapters into a coherent thesis, and the final preparation of the dissertation manuscript were completed by me.

Acknowledgments

I would like to thank the members of my thesis committee—Jan Victor Dee, Xintong Han, Jorgen Hansen, Junnan He, Panos Margaris, Julius Owusu, and Gabrielle Vasey—for their time, support, and constructive feedback. Throughout these acknowledgments, names are listed in alphabetical order by last name.

I also benefited from discussions with colleagues and faculty at Concordia University. Helpful feedback came from audiences at the Canadian Economics Association annual meetings, the Chinese Economists Society annual meetings (North America and China), the Asian Meeting of the Econometric Society, as well as seminars at the University of Macau, Institut Mines-Télécom Business School, Concordia University, and Jinan University. A visiting period at the University of Edinburgh Business School provided additional opportunities for discussion and feedback and helped advance this project.

I am grateful to Christian Belzil, Marie Carpenter, Grazia Cecere, Yanyou Chen, Zhiqi Chen, Prosper Dovonon, Xintong Han, Ming Li, Tongzhe Li, Yushen Li, Xingfei Liu, Damba Lkhagvasuren, Christian Sigouin, Tong Wang, Huan Xie, Jipeng Zhang, and Pu Zhao for their advice, feedback, and many helpful conversations that strengthened this work. I also thank my classmates and friends—Oyu-Erdene Buyandelger, Kefan Chen, Altynai Pankratov, Kyle Phong, Rajrupa Saha, Bilguun Sukhbaatar, and The Anh Vo—for their support on this dissertation, including helpful discussions and feedback. Finally, I sincerely appreciate the support of our departmental administrative staff—Domenica Barreca, Melissa Faisal, Mary-Ann Jirjis, Émilie Martel, and Kelly Routly—for their patience, responsiveness, and help with all the logistics that kept things running smoothly.

This research was supported by a Social Sciences and Humanities Research Council Doctoral Fellowship, funding from Québec's Ministère de l'Éducation et de l'Enseignement supérieur, a Mitacs Globalink Research Award, NET Institute grants, Concordia University's School of Graduate Studies, Faculty of Arts and Science, and Department of Economics, UK Research and Innovation, the University of Edinburgh Business School, and the Government of Alberta. I gratefully acknowledge this support.

Finally, I thank my family for their selfless dedication and boundless support. I also thank my friends for their encouragement and patience.

Table of Contents

List of Figures

List of Tables

Introduction

Over the past two decades, the global economy has undergone a profound transformation driven by digital technologies. The rapid diffusion of digital platforms has reshaped how producers, consumers, and intermediaries interact, giving rise to new modes of exchange, competition, and innovation. From rural e-commerce initiatives that connect farmers with urban consumers to creator platforms that redefine how knowledge and content are monetised, digitalisation has become one of the most powerful forces shaping modern economic and social systems. According to a recent report (Intelligence, 2024), e-commerce sales are expected to exceed 6 trillion dollars in 2024, with digital retail accounting for 20.1% of total global sales. Beyond its commercial scope, digitalisation now sits at the centre of national development strategies, where governments seek to harness technology to foster inclusive growth, bridge information asymmetries, and revitalise lagging sectors.

Despite its global reach, the digital economy has not developed evenly. Many developing regions continue to face structural barriers that limit digital participation, such as low digital literacy, inadequate logistics, and weak brand credibility. These frictions constrain small producers and rural entrepreneurs from fully realising the benefits of online markets. In response, policymakers have increasingly turned to public digital infrastructures—such as government-initiated e-commerce platforms—to reduce entry costs, enhance visibility, and promote inclusive participation. At the same time, the private sector has been at the forefront of digital innovation, developing sophisticated platform architectures and behavioural mechanisms that leverage pricing, social influence, and algorithmic design to sustain user engagement and generate revenue.

This duality between public digital initiatives and private digital ecosystems defines the current frontier of research in digital economics. Both forms of platforms aim to improve efficiency and participation, yet they operate through different mechanisms: public platforms rely on collective action and policy support, while private platforms rely on market incentives and behavioural engagement. Understanding how these two models shape markets and behaviours—and how individuals respond to them—is crucial for assessing the broader welfare implications of digital transformation.

This dissertation investigates these questions through three empirical chapters that together analyse how digital platforms influence both market outcomes and individual decision-making. The first chapter focuses on the production side of the digital economy by studying how a government-initiated e-commerce platform (GEP) affects the sales of a traditional agricultural product in China’s Pu’er tea market. The second and third chapters turn to the demand side, exploring how pricing policies, peer effects, and network structures shape renewal behaviour in a digital content membership platform. Taken together, these studies provide a unified understanding of how digital platforms—public or private—reallocate transactions, reshape incentives, and affect welfare in the emerging digital economy.

The first chapter begins with the question of how digitalisation can revitalise traditional markets. In many developing regions, local agricultural producers have long faced difficulties in reaching wider markets because of high transaction costs, asymmetric information, and limited branding capacity. E-commerce offers the potential to mitigate these frictions by expanding market access and reducing geographic barriers. Yet, the effectiveness of government-led e-commerce programs remains an open empirical question. While governments worldwide have promoted digital participation as a pathway to inclusive growth, it is unclear whether public e-commerce platforms actually stimulate production and sales or merely shift existing transactions across channels.

To address this issue, the first chapter, Digital Revitalization or Useless Effort? Public E-commerce Support and Local Specialty Sales, examines the impact of a government-initiated e-commerce platform (GEP) in China's Pu'er tea market. Using a unique dataset from field experiments and surveys of 983 farmers that cover more than 95% of local tea output over five years, the study investigates how access to the GEP affects online and offline sales. Employing a two-way fixed-effects model, it identifies substantial substitution effects: for tea of a given quality, access to the GEP increases online sales by 16.649% and decreases offline sales by 15.549%, revealing a clear shift from offline to online transactions. On the extensive margin, households that previously sold only offline become more likely to sell online; on the intensive margin, adopters expand their online channels and offer a wider range of tea qualities.

A closer look at adoption patterns reveals that farmers typically enter the online market via social media before moving to e-commerce platforms, including the GEP. This finding suggests that the substitution effect arises not only through the platform itself but also through the GEP's bundled public services, such as cooperative packaging and regional branding. The mediation analysis shows that the increase in online sales channels and product variety explains how GEP access drives the shift toward online transactions. Overall, the chapter demonstrates that public e-commerce programs can facilitate structural transformation within traditional agricultural markets—not necessarily by expanding total output, but by changing how and where transactions occur. These results highlight the potential of government-led digital platforms to foster local economic revitalisation and broaden digital inclusion.

While the first chapter focuses on the production side of digitalisation, the second and third chapters move to the demand side, examining how users behave and make decisions within private digital ecosystems. Digital content platforms—where creators monetise knowledge, ideas, or entertainment through subscriptions—have become a dominant model in the digital economy. The success of these platforms relies not only on pricing and product quality but also on network-based social mechanisms that influence user engagement and retention. Understanding how users decide whether to renew a digital content membership requires analysing both economic incentives and social interactions within referral networks.

The second chapter, Peer Effects and Price Elasticity in Digital Membership Renewal, investigates the behavioural complexity of digital membership renewal on a large Chinese creator platform. Using a rich dataset that traces users' renewal and referral activity, the study examines the interplay of price changes and peer decisions within referral networks and their joint effects on renewal behaviour. Through regression modelling, the chapter quantifies price elasticity and uncovers a positive correlation between a user's likelihood to renew and both the renewal decisions of the referrer and the number of referees. It also documents snowballing effects of price changes that propagate throughout the network, showing that user decisions are influenced not only by direct economic considerations but also by the observed actions of peers.

Building on these empirical findings, the third chapter develops a structural model to endogenise the formation of referral networks and capture users' interdependent renewal decisions. The model allows referral relationships to emerge from users' expectations about the renewal choices of both upstream referrers and downstream referees. This structure enables a more precise identification of network effects and behavioural interdependence than reduced-form models alone. The estimation results indicate that referee-targeted discounts are more effective than either uniform discounts or referrer-targeted discounts in increasing overall renewal rates. Moreover, the model reveals that networks characterised by high connectivity but low centrality are more conducive to sustaining long-term customer loyalty. Counterfactual simulations based on the structural model further demonstrate that small, targeted adjustments in discount strategies or referral structures can lead to significant improvements in renewal outcomes without sacrificing platform revenue.

Taken together, the second and third chapters deepen our understanding of how private digital platforms influence user behaviour and network-based interactions. The reduced-form analysis provides direct empirical evidence of peer influence and price sensitivity, while the structural model captures the endogenous formation of network ties and strategic interdependence among users. These complementary perspectives jointly highlight that successful digital platform management depends not only on pricing policies but also on the architecture of social connections and referral incentives that shape user retention.

The three chapters together address a central question in digital economics: how do digital platforms—whether public or private—transform markets, behaviours, and welfare? The first chapter shows how public digital platforms can overcome structural constraints and reshape producer participation in traditional sectors. The second and third chapters demonstrate how private platforms rely on social networks and behavioural mechanisms to retain users and generate value. By bridging these two perspectives, the dissertation provides a holistic account of the economic and behavioural mechanisms underlying digital transformation.

Collectively, this dissertation contributes to several strands of research in digital economics, development economics, and platform studies. On the supply side, it contributes to the literature on public digital infrastructure, rural e-commerce, and government-led digital transformation by offering causal evidence of how a government-initiated platform affects producers' market participation and channel choice. By combining field experiments with detailed micro-level survey data, the first chapter provides a rare empirical examination of how public e-commerce programs function as instruments of rural revitalisation. It demonstrates that digital participation can be effectively promoted through bundled public services—such as regional branding and cooperative packaging—that complement the technological affordances of the platform itself. This evidence advances our understanding of how government interventions can correct market failures in digital access and reduce inequality in digital participation.

On the demand side, the second and third chapters contribute to the growing literature on digital platform design, user retention, and network-based behaviour. They shed light on the interplay between economic incentives and social influence within referral networks, showing that renewal behaviour on digital content platforms cannot be explained solely by price sensitivity or individual preferences. Instead, renewal decisions are embedded in a social context where peer effects, network topology, and referral structures jointly determine user outcomes. By integrating reduced-form estimation with structural modelling, the two chapters provide both empirical validation and theoretical generalisation of the behavioural foundations of user retention in digital markets. These insights extend the current understanding of platform economics by demonstrating that optimal discount and referral policies depend not only on marginal pricing responses but also on the endogenous organisation of users' social connections.

Taken together, the three chapters reveal a broader picture of the digital economy as a system shaped by both institutional design and behavioural interaction. Public digital platforms reduce participation barriers and create new opportunities for producers, while private platforms depend on behavioural interdependence and social reinforcement to sustain engagement. The dissertation thus bridges two key dimensions of digitalisation—policy-driven inclusion and market-driven retention—illustrating how digital transformation simultaneously empowers producers and reshapes consumer behaviour.

Beyond their specific empirical settings, the findings of this dissertation carry important policy implications. For policymakers, the evidence from the Pu'er tea market suggests that the success of digital public programs depends not only on technological access but also on the provision of complementary services that build trust and brand recognition. For platform managers, the analysis of referral networks highlights the importance of targeting behavioural incentives efficiently—designing discount strategies and social mechanisms that encourage renewal while preserving revenue. More broadly, the results underscore the need for policies and business strategies that balance inclusivity, efficiency, and sustainability in the digital economy.

The remainder of this dissertation is structured as follows. Chapter 1 presents the study *Digital Revitalization or Useless Effort? Public E-commerce Support and Local Specialty Sales*, which analyzes the causal effects of a government-initiated e-commerce platform on farmers' online and offline sales. Chapter 2 examines *Referral Networks and Renewal Rates: An Empirical Study from a Creator Platform*, using a reduced-form empirical analysis to quantify the effects of price changes and peer decisions within referral networks. Chapter 3 develops a structural model of network formation and renewal behaviour, enabling counterfactual simulations of alternative referral and pricing strategies. The dissertation concludes with a synthesis of the main findings, a discussion of their theoretical and policy implications, and suggestions for future research in digital economics and platform governance.

Chapter 1: Digital Revitalization or Useless Effort? Public E-commerce Support and Local Specialty Sales

The distinct advantages of e-commerce over traditional channels have led to an unprecedented surge in its adoption across many regions. Specifically, the shift to online transactions would remove fixed entry costs and eliminate geographic barriers to trade. According to a recent report (Intelligence, 2024), e-commerce sales are expected to exceed 6 trillion dollars in 2024, with digital retail making up 20.1% of total sales. As e-commerce platforms become more popular and widely used, policymakers are increasingly recognizing their transformative potential as a driver of agricultural productivity and rural economic growth, especially in low and middle-income countries (LMIC). In India, the government launched its own Open Network for Digital Commerce (ONDC) in 2022 and has been promoting it nationwide. Designed to compete with Amazon, this platform offers a variety of products, including groceries, beverages, and other consumer goods (Mandavia, 2022). Similarly, China's e-commerce policy has been a major focus in recent decades, guided by central government directives. Despite government initiatives to facilitate rural e-commerce, small farmers continue to face significant barriers. These include logistical challenges, a lack of expertise, and difficulty building trust and brand recognition. In response, the Chinese government launched public e-commerce support programs to help farmers overcome these obstacles (World Bank, 2019; Vidal and Faz, 2020).

Starting in 2014, the Chinese central government launched the National Rural E-commerce Comprehensive Demonstration Program to support rural counties in developing e-commerce centers (Ma et al., 2023; Li et al., 2025). Beginning in 2017, many local governments across China established government-initiated e-commerce platforms (GEPs) in their regions.[1] Despite the proliferation of GEP programs, empirical evidence on their effects remains scarce, particularly on farmer online versus offline sales choices and the ensuing economic impacts. Furthermore, it is unclear whether rural GEPs and their bundled public services make a meaningful contribution to local revitalization. Progress has been limited by two obstacles: (i) exogenous policy variation suitable for causal identification is rare, and (ii) the collection of microlevel data on farmer output, channel use, and sales volumes requires extensive fieldwork.[2]

Drawing on detailed micro-level data, we analyze the impact of local GEP access on local farmers' sales channel decisions in Pu'er City, Yunnan Province. Our region of study serves as China's primary Pu'er tea cultivation hub: farmers in this region produce about 90% of all Pu'er tea. Historically, these farmers sold their tea in bulk to tourists or large tea processing factories.

---

[1] In Appendix A, we provide a table that lists several GEPs established by local governments in China from 2017 to 2023.

[2] Couture et al. (2021) is an exception. They use household-level data in a randomized controlled trial (RCT) to analyze the impact of rolling out a commercial e-commerce program nationally, and find substantial benefits for rural households (e.g., lower living costs). However, their study mainly examines the entry of the largest Chinese e-commerce platform into rural areas, where most farmers were buyers rather than sellers on those platforms. In contrast, our paper focuses on producers' sales channel choices and examines how a government-initiated support platform reshapes the allocation of sales between online and offline channels. Our setting also allows us to study the role of bundled public services—such as training, packaging, and regional branding—in facilitating producer participation in online markets.

Local authorities responded to the central government's e-commerce initiative by launching a GEP in 2018, allowing farmers to sell directly to consumers. When an order is placed on the platform, farmers process their tea leaves into cakes at a very low cost in government-established cooperatives. The cooperatives label the cakes and ship them to customers. More broadly, treatment in this paper refers to access to the GEP program and its associated public support package, rather than use of the storefront alone. In practice, these public services—such as training, cooperative processing, packaging, and regional branding—can also lower barriers to online selling through other channels, including social media and private platforms.

One of the main empirical challenges of this study is to distinguish the effects on online versus offline sales (Johnson et al., 2017). Before the launch of GEP, very few farmers sold tea online. Even after the platform became available, most tea transactions were still conducted offline. Over a two-year period, we conducted a household survey in six regions of the two main tea-producing areas in Pu'er City. Importantly, these regions adopted the local GEP in a staggered fashion between 2018 and 2020. With the help of the local government, we organized six survey teams, each led by a village leader or a local tea expert, to carry out the data collection.[3] Our data set covers more than 95% of local tea farmers, capturing information on production such as the number of trees and the yearly output of different quality teas. Most importantly, for each year, it records the quantities each household sold through offline and online channels. Therefore, our data set offers a unique opportunity to causally examine the effect of GEP access on farmers' specialty sales and the underlying economic mechanisms.

We employ a two-way fixed-effects (TWFE) regression model, including fixed effects per household and year to control for unobservable household factors and common time trends, to evaluate the causal impact of the GEP on online sales. To address potential within-village correlations over time, we cluster standard errors at the regional level. Our model accounts for variation in the number of nearby tea-processing factories and shipping companies as proxies for changes in the local market infrastructure and access to processing services. This enables us to analyze how these factors affect farmers' online channel choices over time. The results indicate a significant substitution effect: for tea of a given quality, the offline tea sales decrease by 15.549% on average, while online sales increase by 16.649%. Given recent critiques of TWFE regressions (De Chaisemartin and d'Haultfoeuille, 2020; Jakiela, 2021; Roth et al., 2023), we use the interaction-weighted estimator of Sun and Abraham (2021) to address the potential biases that could arise from staggered adoption. We also perform a number of robustness checks, including placebo tests and additional diagnostics to confirm the reliability of our results. These tests confirm that the observed substitution effects originate from the introduction of GEP and are not driven by unobserved factors.

To investigate the mechanisms behind the substitution effect, we stratify households by characteristics measured before GEP access, including annual output, cultivated area, and quality of tea. Total sales remain roughly unchanged across groups, but there is a clear and statistically

[3] Each team included college and university students on vacation, as well as young, educated local residents. The general manager was responsible for coordinating the survey work and cleaning the data.

significant shift from offline to online, with the strongest impact for larger farmers who sell premium-quality tea. We also find both extensive and intensive margin responses: previously offline producers start selling online, and producers who were already selling online expand their use of online channels and increase their product offerings. Most farmers use social media as their initial online channel rather than the GEP, especially first-time adopters. This pattern suggests spillovers from complementary public services surrounding the GEP, including training, cooperative packaging, and regional branding, which lower capability and credibility barriers, allowing farmers to benefit even without utilizing the GEP storefront. For sellers already online with premium tea, the program provides a low-cost route to list lower-end tea, thereby increasing the variety available online. Additional evidence shows that once channel breadth and product variety are included as mediators, the substitution effect attenuates to near zero.

1. Relevant Literature and Contributions

Our analysis of a government-initiated e-commerce support program leverages household panel data and its staggered rollout. This approach enables us to distinguish between extensive-margin effects (i.e., who starts selling online) and intensive-margin effects across sales channels and tea quality levels.

1.1 Market Access and Transaction Volumes

In LMICs, reducing barriers to information and transportation can expand market access and transaction volumes. For example, studies on the adoption of mobile phones have shown that improved communication technology reduces price variation and waste in remote markets, thus improving efficiency in fishing and agriculture (Jensen, 2007; Aker, 2010). Similarly, new transportation infrastructure has broadened the reach of the market: for example, the expansion of railroads in colonial India significantly increased interregional trade and real income (Donaldson, 2018). Likewise, investments in rural road connectivity have been shown to reduce poverty by connecting remote producers with markets (Aggarwal, 2018). Increased Internet access has further improved market integration. Even before the rise of dominant e-commerce platforms, the spread of the Internet in China led to notable increases in export growth for local firms, as it improved communication (Fernandes et al., 2019). Consistent with this, recent evidence from Africa suggests that expanding mobile broadband coverage in rural areas increases household consumption and reduces poverty rates, primarily by enhancing market opportunities and labor outcomes for previously isolated communities (Bahia et al., 2024). Regarding rural e-commerce, direct-to-farmer procurement has raised farmer sales prices and output in India (Goyal, 2010), while the randomized expansion of e-commerce in rural China has resulted in significant gains in consumer welfare but only modest changes in producer income (Couture et al., 2021). In terms of platform design, studies show that ensuring seller reputations (through ratings or certification) and providing buyer protections can mitigate adverse selection in online markets (Jin and Kato, 2006; Saeedi, 2019). Consistent with these findings, the commercial platform rules (fees, deposits, etc.) differ sharply from those of the commission-free GEP (see Appendix B). This contrast implies that lowering fixed and operational barriers (as the GEP does) should shift transactions toward the new platform.

Public e-commerce support programs differ from purely private e-commerce entry. By bundling services such as training, cooperative processing, and regional branding, the GEP in our study serves as an enabling infrastructure for small producers, rather than competing with them as an online platform. Our study expands the literature by examining a government-initiated e-commerce marketplace in a rural setting. We explore how improved market access through this platform affects the distribution of transactions across both quality levels and sales channels. Using the staggered rollout of the platform, we isolate extensive margin effects (which producers begin selling online) from intensive margin effects (the range of products listed online). In doing so, we uncover reallocation patterns that would be concealed in more aggregated data. Our producer-focused findings complement evidence on consumer benefits from commercial platforms (Couture et al., 2021) and on how digital connectivity opens market opportunities (Bahia et al., 2024; Fernandes et al., 2019). This study provides new micro-level evidence from the producers' perspective: public digital support services can reallocate sales from offline to online and expand the range of products sold online, despite short-run supply constraints.

1.2 Supply-Side Reallocation and Market Dynamics

The entry of new sales channels can shift market shares and change the product mixes of firms. Empirical research shows evidence of both cannibalization (new channels taking share from existing ones) and expansion (market growth) effects when a new sales channel enters. For example, the introduction of bike rental services increased total ridership, even as existing bike rental companies lost some customers (Cao et al., 2021). Competition between online food delivery platforms boosted overall usage and revenue, mostly benefiting higher-quality incumbents, rather than simply splitting the existing market (Reshef, 2023). On the supply side, online markets enable a wide range of niche or unique products to find buyers. Increasing variety can attract new consumption rather than simply replacing existing ones (Brynjolfsson et al., 2003). In LMICs, gaining access to larger markets often encourages producers to diversify their offerings or improve the quality of the product rather than simply increasing the quantity. For example, when small rug manufacturers in Egypt were randomly given access to high-income foreign buyers, they responded by significantly improving product quality while reducing the output per hour, rather than increasing total production (Atkin et al., 2017).

In our context, we document a supply-side reallocation mechanism in which the GEP encourages farmers to sell online and diversify their product range. Farmers who adopt early expand the variety of tea products they sell online, including lower-end teas with limited local demand. We observe heterogeneous effects: larger farmers with higher output shift a greater share of their sales online and introduce more product varieties after the GEP, consistent with evidence that better-endowed producers tend to adopt more profitable innovations earlier and more extensively (Foster and Rosenzweig, 1995). From a general equilibrium perspective, our findings help explain why total sales do not increase despite the introduction of the new channel. With fixed short-term production capacity, gains in online sales are offset by decreases in offline sales (Foster and Rosenzweig, 2004). This substitution effect aligns with the view that supply constraints can limit the impact of market expansion efforts. This finding aligns with other evidence from developing markets: in Kenya, an experimental increase in grain traders had a minimal impact on prices or quantities, largely due to high entry costs and persistent market power among intermediaries (Bergquist and Dinerstein, 2020). Therefore, without complementary investments to ease production constraints, digital integration can alter who is reached and what is sold rather than increasing overall production. Consistent with this literature, we find evidence that the public e-commerce platform and its associated public services increased online participation and assortment, rather than total sales.

1.3 Institutional Role and Design of GEPs

Digital platform markets often feature strong network effects and distinct competitive behaviours. A long-standing question is whether market forces result in a single dominant platform or allow multiple platforms to coexist. Classic theory suggests that network externalities can push markets towards a single dominant platform as the optimal outcome from both social and private perspectives (Katz and Shapiro, 1994; Shapiro and Varian, 1999). However, subsequent research has challenged this view, showing that smaller platforms can survive by differentiating themselves and targeting niche markets rather than competing in a winner-take-all scenario (Cennamo and Santalo, 2013). Empirical evidence from studies of new platform entries in established markets reveals a variety of outcomes. For example, the emergence of a peer-to-peer rental platform such as Airbnb greatly cannibalized the business of existing hotels in affected areas (Zervas et al., 2017). Similarly, the launch of ride-sharing services such as Uber caused significant declines in earnings for traditional taxi drivers (Berger et al., 2018). These cases demonstrate that privately operated digital platforms often act as new competitors for incumbents, eroding existing market shares.

We consider the public-service aspect as an essential part of the impact of the GEP: The range of government-provided services (including training, cooperative processing, and branding support) is integrated into the platform's operations and helps producers who were previously excluded from participating in e-commerce. Therefore, the objective of our study is to estimate the effect of a public-service ecosystem, rather than that of a government-initiated marketplace competing with giant commercial platforms (e.g., Taobao or JD). Our micro-level findings contribute to the broader development literature on the role of government in market integration, suggesting that for small-scale producers in LMICs, public digital services complement private platforms by reducing both financial and operational costs. For example, policy reforms that improved local governance and reduced bureaucratic steps in China have been shown to boost economic performance by enhancing market efficiency (Li et al., 2016). Similarly, special economic zones created by the government have significantly increased output and exports in Chinese cities by easing business restrictions (Wang, 2013). Likewise, the GEP functions as an infrastructural intervention that reduces barriers to entry for rural sellers. Although we do not have direct evidence on the long-term viability of the GEP itself, we document that the public service program increases the multichannel presence of households. This pattern suggests that the introduction of GEP generates spillovers, helping farmers engage in online sales through various channels beyond the platform's own site.

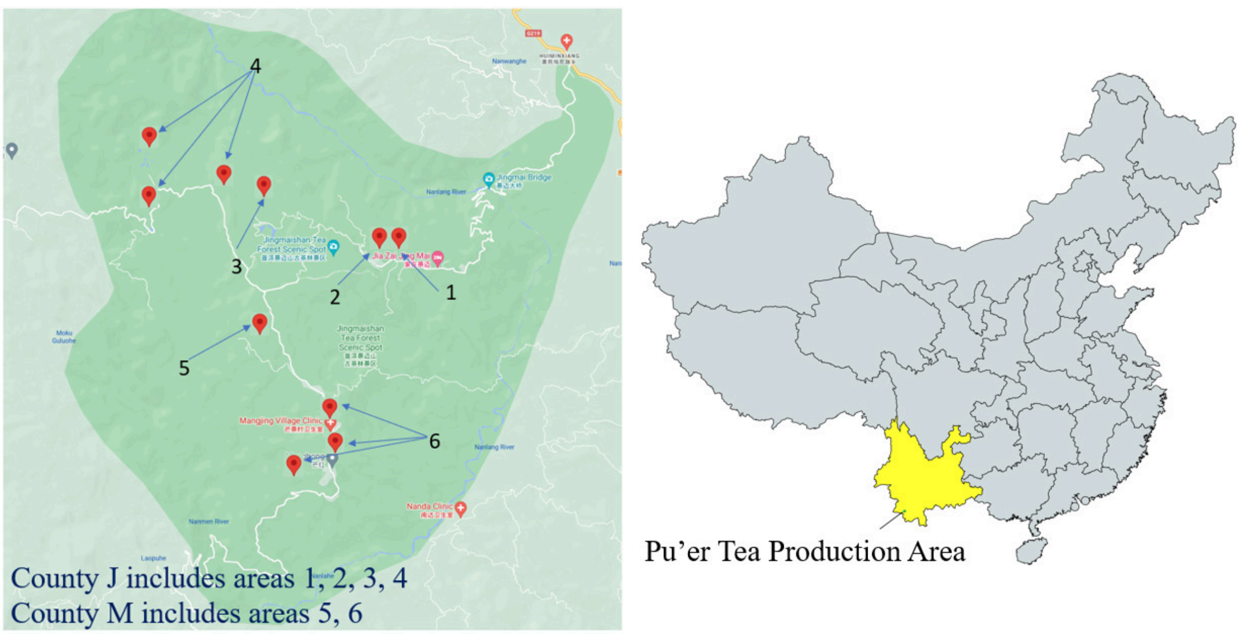
2. Institutional Background

This section provides an overview of Pu'er tea production and sales. It explains the central government's guidelines for promoting online Pu'er tea sales and analyzes how local governments adapt these policies to suit local needs.

2.1 A Brief Description of the Local Specialty

This study focuses on Pu’er tea—a unique variety grown predominantly in six mountainous areas around Pu’er City, Yunnan Province. In 2008, Pu’er tea received protected Geographical Indication status from the Chinese government, which legally restricts the use of the name “Pu’er” to tea produced in that region. The study focuses on two major tea farming counties in Pu’er City (Counties J and M), home to the world’s largest ancient tea forest. Figure 1 shows the location of Yunnan within China and highlights the main Pu’er tea zones. It also marks the exact locations of two counties, along with the six areas surveyed. Data for this study are mainly obtained from surveys conducted with residents of these areas.

The data set includes farming output and sales data for three primary tea varieties in the region: premium-quality, high-quality, and regular tea, which together make up over 98% of the local tea farming output during the study period. Premium-quality tea is characterized by its origin from the central buds of single, ancient trees over 50 years old, representing the highest grade. High-quality tea is a mixture of leaves from several ancient trees, each over 20 years old. Regular tea, the most commonly produced type, is harvested in spring from younger trees in plantation tea gardens.



Notes: The map on the left shows the six surveyed areas in Pu’er City, Yunnan Province, which are grouped into two administrative counties: County J (Areas 1–4) and County M (Areas 5–6). These areas are located within the world’s largest ancient tea forest, recognized as a UNESCO World Heritage Site in 2023. The map on the right highlights Yunnan Province (in yellow) within mainland China. Pu’er tea is a protected geographical indication tied to designated areas in Yunnan. The depiction of map boundaries is for research illustration only and does not imply any position by the authors or the publisher on jurisdictional claims.

Figure 1. Geographical Location of Pu’er Tea Farming Areas

2.2 Policy Treatment: Launch of the Platform

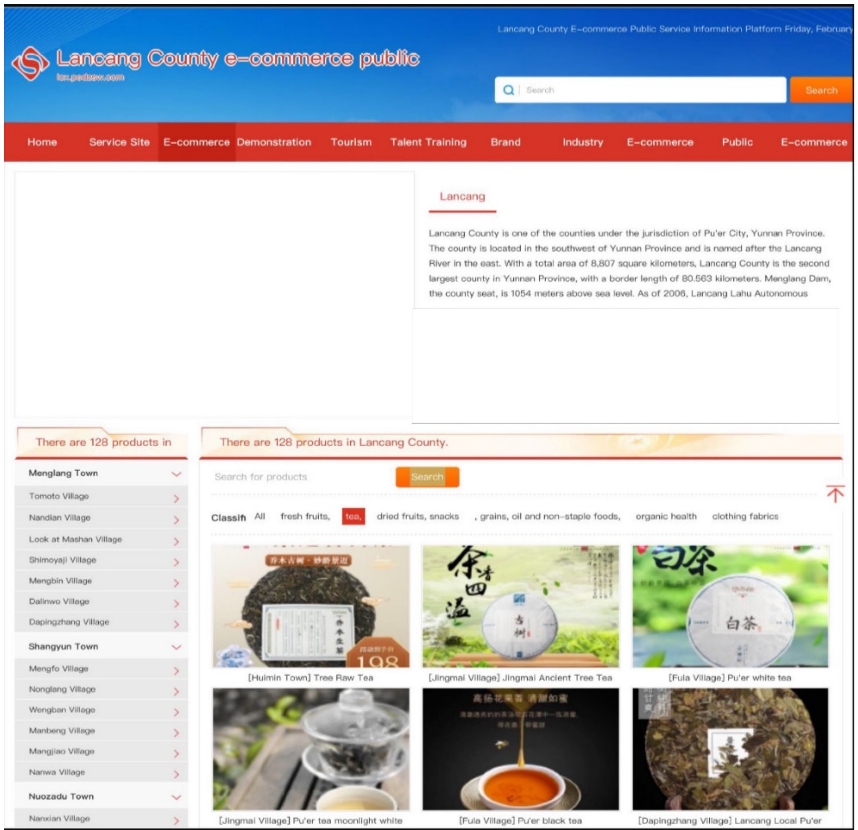
The Chinese government has initiated a series of efforts to promote e-commerce for specialty products in rural areas. In 2017, the Ministry of Commerce and the Ministry of Agriculture and Rural Affairs jointly issued a directive (Commerce Construction Letter [2017] No.597) calling for "deepening agro-commercial cooperation and vigorously developing agricultural e-commerce." This and related policies encouraged local governments to build an infrastructure of public services in e-commerce in rural regions, for example, by establishing government-supported e-commerce public service centers and training programs that help farmers sell their agricultural products online (Ma et al., 2023; Li et al., 2025). Starting in 2014 (and expanded in 2017), the Central Government's National Rural E-Commerce Comprehensive Demonstration Program allocated 20 million RMB per pilot county to develop e-commerce support systems (Ma et al., 2023; Li et al., 2025). The aim was not to build entirely new online marketplaces, but to improve the local e-commerce environment, offer training services, and develop regional brands as public goods.

In response to the central government initiative, the local government launched the Lancang County E-commerce Public Platform in late 2017 (with full implementation in early 2018) as a support system for online tea sales from farmers. This platform (essentially a regional e-commerce service center) was created to guide farmers in marketing their tea through digital channels. In practice, the Lancang platform is government-initiated and commission-free, operating as a public service rather than a profit-driven platform. Through the platform arrangements, farmers receive online orders from external customers and process raw tea leaves into compressed tea cakes at local cooperatives to fulfill orders. The cooperatives, established by the local government, provide packaging and branding services to farmers at a minimal cost (5 RMB per kilogram of processed tea). The finished tea cakes are labeled with the cooperatives' regional brand, lending credibility and a shared regional identity to the product (Yunnan Provincial Department of Natural Resources, 2024). This branding support through cooperatives and regional labels helps small farmers overcome trust and recognition barriers in online markets. The minimal processing fees and zero commissions enable farmers to engage in online sales without incurring the high costs typically imposed by commercial e-commerce platforms.

The platform was introduced at the end of 2017, with implementation starting in early 2018 across various regions, creating a valuable quasi-experimental setting for our analysis.[4] Figure 2 shows a snapshot of the platform's website interface. Importantly, the Lancang local government also invested in training and outreach programs to ensure farmers could effectively use this new channel. During the initial rollout, government personnel were dispatched to villages throughout the county to introduce the platform to tea farmers and demonstrate how to participate, including registering and listing products, adhering to online quality standards, and managing online orders. All tea producers in the area were informed about the platform and encouraged to sell their tea through it. Additionally, the government organized interactive training sessions via

[4] In Appendix C, we document the timeline of adoption of the platform in various townships and villages, highlighting the staggered rollout and its timing in each location.

social networks, establishing WeChat groups that included local farmers and platform administrators. These groups served as continuous support and compliance training forums where farmers could ask questions, share experiences, and receive timely guidance on online sales.



Notes: Screenshot of the Lancang county portal (translated). The interface allows filtering by region (lower left) and lists cooperative-branded tea products (bottom).

Figure 2. Snapshot of Lancang County E-commerce Public Platform

Lancang's government was not alone in this effort - in fact, many local governments throughout China established similar public e-commerce service programs in the late 2010s in line with the national demonstration policy. In Appendix A, we list public service centers for e-commerce, established by various counties, alongside similar platforms, from 2017 to 2023, as part of this rural digitization initiative. Appendix B further provides more details and a comparative discussion of Lancang's public e-commerce service versus other online sales channels available to farmers. Despite the nationwide rollout of such public e-commerce centers, their effectiveness in boosting rural incomes remains largely understudied. Our study addresses this gap by carefully evaluating Lancang's program at the household level, analyzing how access to the GEP affected sales patterns for tea farmers in the region.

3. Data Collection and Descriptive Statistics

This section describes the two administrative counties in the sample, describes the survey methodology, and explains the steps taken to ensure data quality. The second half of this section presents the descriptive statistics.

3.1 Data Collection

To gather data on farmers' output, sales volumes, and sales channels, we selected two administrative counties (pseudonymously County J and County M) in the Pu'er tea-growing region for our household survey. The locations of both counties and their subareas are shown in Figure 1. County J comprises four distinct areas covering a total of 66.9 square kilometers. Its population is 3,339 people living in 801 households. A large portion of the population has lived in this mountainous region for many generations. To the south of County J is County M, which comprises two areas and has a population of 2,645 people, spread across 639 households.

A household survey was conducted in Counties M and J to collect the data needed for the study. A comprehensive survey was carried out in each area. The survey team was organized in collaboration with the local government. In County J, 785 of 801 households (98%) provided complete questionnaire responses. In County M, 198 households were surveyed and responded to the questionnaire. It is important to note that this does not imply that County M had a lower response rate. Unlike County J, where tea farming is nearly universal among households, County M has only 202 households involved in tea farming. Together, our survey covers more than 98% of tea-farming households in these areas, producing an almost complete household panel for the years 2016–2020. For each household, data were collected on variables such as the annual production of tea, the yearly area of accessible agricultural land, the yearly output of different quality tea leaves, and the quantities sold online and through offline channels. In Appendix D, we provide a detailed description of the data collection method and show a photo of the interview process.

3.2 Summary Statistics

Table 1 summarizes the key variables in our dataset. Tea sales occur through two primary channels: offline and online. Offline, tea farmers typically sell raw leaves to nearby processing factories right after harvest. By contrast, the online sales avenues consist of two types: social media (e.g., TikTok or Douyin) and e-commerce platforms (e.g., Taobao and Tmall). Before 2018, most farmers who sold online did so through social networks because they had lower entry and operating costs compared to e-commerce platforms, which often required brand registration and relevant certifications. When we consider 2018 as the year of policy change and overlook the specific adoption times for farmers in different areas accessing the GEP, it becomes clear that online sales of the three grades of tea (i.e., premium-quality, high-quality, and regular tea) have increased significantly after 2018. In contrast, the average offline sales figures for all grades of tea decreased compared to the pretreatment period. Despite these notable substitution effects, online sales still account for less than half of total household sales, as of 2018.

Table 1. Summary Statistics for Sales Volume

	Before 2018			After 2018		
	Premium	High	Regular	Premium	High	Regular

Sales Volume (Kg)	Online	113.50	104.33	176.67	199.54	188.91	375.01
		(79.34)	(75.53)	(181.23)	(129.51)	(120.29)	(409.94)
	Offline	459.84	391.49	873.37	394.04	346.62	832.72
		(279.23)	(257.69)	(725.81)	(264.01)	(235.28)	(691.25)

Notes: We report the standard deviation in parentheses.

In addition to the variables listed in Table 1, our sample also includes detailed information on time-varying household items in these six areas. Additional data have been collected on the number of tea processing factories and freight companies located near each area during the five-year period covered by the dataset. In Appendix E, we provide a further statistical summary of the data, broken down by household and area. We find that the sizes of agricultural land for each farmer, as well as the number of factories and shipping companies, remain relatively stable during our sample period. This suggests that the local environment did not undergo significant changes other than the introduction of the GEP, and that the shift toward online sales is not primarily driven by these factors.

4. Econometric Model and Estimation Results

In this section, we introduce the econometric model used to assess the impact of GEP on tea sales and discuss its key identification assumptions. We then present the baseline results and conclude with a series of robustness checks employing alternative estimators and specifications.

4.1 Econometric Model

To quantify the effect of launching the GEP, we employ an econometric model with the following specifications:

$$q_{i,j,t} = \alpha + \gamma D_{i,t} + \delta mode_{i,j,t} + \theta D_{i,t} \times mode_{i,j,t} + \zeta \times Z_{i,j,t} + \mu_i + \eta_j + \psi_t + \epsilon_{i,j,t}, \quad (1)$$

where $q_{i,j,t}$ denotes the logarithm of the total amount (in kg) of quality j leaves sold by household i in period t. Accordingly, the unit of observation in the regression is a household–quality–period–channel observation rather than a household alone. Since the data cover 983 households, 3 tea-quality categories, 5 years, and 2 sales channels, the estimation sample contains about 29,490 observations. $D_{i,t}$ represents the treatment variable. It equals 1 if household i's area has gained access to the GEP in period t, and 0 otherwise. The $mode_{i,j,t}$ is a binary indicator that captures online and offline sales channels. This variable is equal to 1 if the sale of j quality tea leaves by household i in period t was made through an online channel.

The use of panel data enables the incorporation of two-way fixed effects, specifically household- and year-fixed effects. The potential impact of time-invariant unobserved household characteristics and time trends on estimates is captured by variables μ_i and ψ_t, respectively. Additionally, a binary variable, $Z_{i,j,t}$, is included, taking a value of 1 when the farming output of type j tea leaves by household i in period t is 0. This approach helps avoid biased estimates that could result from some households producing only a single type of tea. The variable ζ captures this effect. The variable η_j indicates fixed effects at the quality level. The household level time-varying unobserved quality-specific error term is represented by the variable $\epsilon_{i,j,t}$.

Equation 1 closely resembles a traditional difference-in-differences econometric model. The model captures the first difference by comparing tea sales at the household level before and after access to the GEP, designated as the treatment variable $D_{i,t}$. The second difference is obtained by the mode of sale. That is, the change in online sales before and after treatment is compared to the change in offline sales before and after the treatment. By calculating these differences, we can estimate the effect of treatment on offline and online sales through the parameters γ and θ as long as $D_{i,t}$ is assigned randomly.

4.2 Baseline Results

Table 2 presents the baseline results. Column (1) shows estimates with no fixed effects, Column (2) adds year fixed effects, and Column (3) also includes both household and quality fixed effects. The estimated coefficients are statistically significant at the 1% level, indicating that the policy has a significant impact on how households choose between online and offline sales channels. The following is a description of the interpretation of our coefficients. Following the acquisition of access to the platform, the volume of offline sales is observed to increase or decrease by an average of $100 \times (exp(\gamma) - 1)\%$. Similarly, online sales experience an average increase or decrease of $100 \times (exp(\gamma + \theta) - 1)\%$ after obtaining access to the platform. Based on Column (3) of Table 2, after an area gains access to the GEP, online sales increase by 16.649% on average for tea of a given quality. In contrast, offline sales drop by an average of 15.549% after the area has access to the GEP. Our findings indicate a statistically and economically significant shift by households from selling their tea through offline channels, such as factories and local markets, to online channels, including the GEP and various social media and commercial e-commerce platforms. The relatively high R^2 mainly reflects the structure of the data and the inclusion of the zero-output indicator, which absorbs substantial variation in household–quality–year sales, many of which are structurally close to zero.

Table 2. Effect of the GEP Access on Sales

Dependent Variable:	Log(sales): $q_{i,j,t}$					
	Without Clustering			**With Clustering**		
	(1)	(2)	(3)	(4)	(5)	(6)
Platform Access (γ)	−0.148***	−0.179***	−0.169***	−0.148**	−0.179**	−0.169***
	(0.011)	(0.013)	(0.014)	(0.049)	(0.065)	(0.035)
Platform Access × Online Sales (θ)	0.316***	0.316***	0.323***	0.316***	0.316***	0.323***
	(0.015)	(0.015)	(0.014)	(0.076)	(0.076)	(0.078)
Online Sales (δ)	−0.474***	−0.474***	−0.482***	−0.474***	−0.474***	−0.482***
	(0.008)	(0.008)	(0.007)	(0.036)	(0.036)	(0.038)
Zero Output (ζ)	−5.484***	−5.483***	−5.429***	−5.484***	−5.483***	−5.429***
	(0.007)	(0.007)	(0.007)	(0.073)	(0.072)	(0.065)
Constant (α)	5.739***	5.748***	5.715***	5.739***	5.748***	5.715***
	(0.007)	(0.007)	(0.007)	(0.094)	(0.096)	(0.055)
Observations	29,490	29,490	29,490	29,490	29,490	29,490

Quality FE	NO	NO	YES	NO	NO	YES
Household FE	NO	NO	YES	NO	NO	YES
Year FE	NO	YES	YES	NO	YES	YES
R^2	0.956	0.956	0.965	0.956	0.956	0.965

Notes: Standard errors are indicated in parentheses. In Columns (4), (5), and (6), error terms are clustered at the area level. Significance levels are denoted as follows: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

Although panel data helps us better isolate the treatment effect by controlling for various fixed effects, we remain cautious about potential correlations in the error term caused by unobserved area-specific factors that change over time. According to our baseline specifications, the unit of analysis is household-quality-year sales (both online and offline). However, our treatment, which is gaining access to the GEP, occurs at the area level rather than the individual level. As a result, we cluster standard errors at the area level for our baseline model and all subsequent models where this discrepancy exists. Estimates for the baseline specification, with standard errors clustered at the area level, are shown in Columns (4), (5), and (6) of Table 2. After clustering standard errors at the area level, our estimated treatment effect remains statistically significant at the 1% level.

Overall, these estimates should be interpreted as intent-to-treat effects of gaining access to the GEP at the area level, rather than as the effect of household-level take-up of the government storefront. This is because treatment is defined by area-level access to the GEP, and some treated households do not directly use the platform even after access becomes available.

4.3 Robustness Check 1: Unobserved Trends and Environmental Changes

One of the key assumptions underlying the identification of our model is that no factors changing over time at the area level are correlated with both treatment (access to the GEP) and result (tea sales). For example, imagine a scenario where farmers become more productive over time. In that case, an incorrect conclusion might be drawn, attributing the increase in online sales to the introduction of the GEP. However, the increase in online sales could actually be due to a boost in farming output. Therefore, it is crucial to address these issues to accurately measure the impact of GEP on online and offline sales. To achieve this, we incorporate additional control variables into our model, including the volume of tea produced by each household and the number of factories and shipping companies in each area. These controls help us distinguish the effects of the policy from other area-specific factors that change over time and could affect our outcome variables. We estimate the following equation:

$$q_{i,j,t} = \alpha + \gamma D_{i,t} + \delta mode_{i,j,t} + \theta D_{i,t} \times mode_{i,j,t} + \zeta \times Z_{i,j,t} + \beta' X_{i,t} + \mu_i + \eta_j + \psi_t + \epsilon_{i,j,t}, \quad (2)$$

where $X_{i,t}$ is a vector of controls that includes the log of the amount of tea produced by household i in year t and the number of factories and shipping companies in the area of household i in year t.

Furthermore, even after including area-level controls, we acknowledge the possibility of unobserved area-specific, time-varying factors that could be linked to both the treatment and outcome variables. For example, if certain areas adopt smartphone technology more quickly than others, those with higher smartphone adoption rates may show higher online sales. Ignoring these unobserved time-varying differences across areas could lead to biased estimates of the effect of treatment. To address this, we incorporate area-specific trends into Equation 2. Specifically, we estimate the following equation:

$$q_{i,j,t} = \alpha + \gamma D_{i,t} + \delta mode_{i,j,t} + \theta D_{i,t} \times mode_{i,j,t} + \zeta \times Z_{i,j,t} +$$

$$\beta' X_{i,t} + \mu_i + \eta_j + \psi_t + area_i \times TT_t + \epsilon_{i,j,t}, \tag{3}$$

where $area_i$ represents the area where household i resides. TT_t denotes a polynomial time trend, and $area_i \times TT_t$ constitutes the interaction term.

Detailed results are shown in Appendix F. The first two columns of Table F.1 indicate that online sales increase by an average of 18.412%, while offline sales decrease by an average of 16.222% when controls for area-specific factors varying over time are included in Equation 2. The last two columns incorporate additional controls for household-level farming output and area characteristics in Equation 3, respectively. Our results are consistent with different specifications, suggesting that there are no significant trend differences across counties.

4.4 Robustness Check 2: Treatment Endogeneity

The presence of unobserved variables that influence both treatment and outcome simultaneously can also cause endogeneity bias in our estimates. This bias may mistakenly assign the effects of these hidden factors to the treatment itself (Angrist and Pischke, 2009). To improve the causal interpretation of our estimates, it is crucial to confirm that the timing of the GEP access is not associated with unobserved time-varying factors at the area level that could also affect tea sales. While the local government confirmed that no specific criteria or special considerations were used to determine which area first gained access to the platform, we conducted additional tests to verify randomization.

The first step is to determine whether the probability of gaining access to the GEP depends on area-level characteristics, such as the level of farming output, the number of tea processing factories, and the number of shipping companies. Following the methodology proposed by Zervas et al. (2017), we analyze whether the GEP access was systematically related to these area characteristics. The detailed results are provided in Appendix G. Our results indicate that area-specific time-varying factors, such as total tea production, the number of factories, and the number of shipping companies, are not correlated with the timing of GEP adoption.

To further strengthen the robustness of our findings, we propose a series of placebo tests. These tests aim to verify whether the estimated effect of treatment truly reflects the impact of the government's policy intervention, rather than being influenced by other confounders related to treatment and the farmer's choice of online or offline sales channels.

In our first placebo test, we randomly assigned the years during which a household or area has access to the platform, while keeping the total number of years of access fixed. The results of this test are shown in Columns (1) and (2) of Table G.2 in Appendix G. In Column (1), we reshuffle treatment at the area level. For example, if an area had access to the GEP in 2019 and 2020 (a two-year period), we randomly select two years between 2016 and 2020 and assign a value of one to a new variable called ‘placebo treatment’ for those selected years. The placebo treatment is applied consistently to all households within a given area. In Column (2), the treatment status is reshuffled for each household, rather than each area. After creating the placebo treatment, we then estimate its effect on offline and online sales. The results of both columns show that placebo treatment does not have a statistically significant effect on online or offline sales of a household at the significance level 10%. In the second placebo test, we estimate Equation 1 using a subset of households that have never participated in online sales during the entire sample period. Our data indicate that roughly 9% of the total sample falls into this category. If the impact of the GEP on tea sales across different channels is solely due to the launch of the GEP, these non-online sellers should remain unaffected by the policy change. The results, presented in Column (3) of Table G.2 in Appendix G, support this idea. We find that the GEP had no effect on sales volumes for non-adopters.

4.5 Robustness Check 3: Bias Correction Related to TWFE Estimators

As highlighted by recent econometric studies (De Chaisemartin and d’Haultfoeuille, 2020; Jakiela, 2021), the estimation of TWFE is unbiased when the effects are homogeneous across units and periods. In other words, when there are no dynamic changes in the effects of the treatment. The bias in TWFE estimation persists even when treatment is randomly assigned, as interactions between treatment effects and time still occur with random assignment. This section includes additional robustness checks to prevent bias that arises when previously treated observations are implicitly used as controls for newly treated observations.

4.5.1 Negative Treatment Weights

Following Jakiela (2021), $\hat{\theta}^{TWFE}$ in Equation 1 can be derived using the Frisch-Waugh-Lovell Theorem:

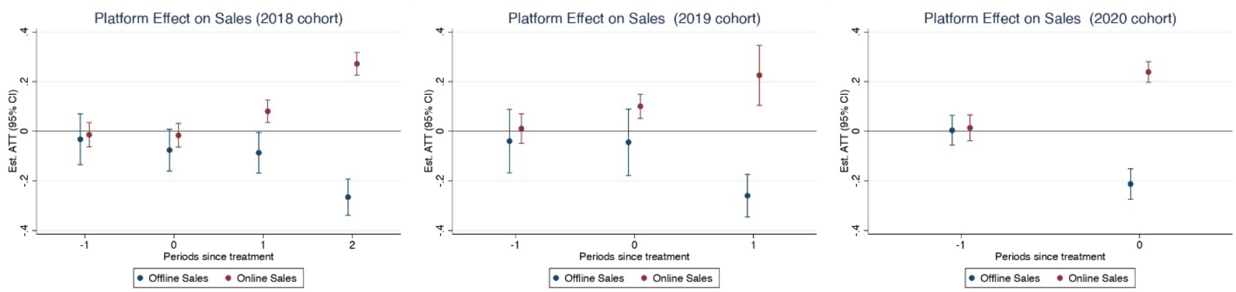
$$\hat{\theta}^{TWFE} = \sum_{ijt} q_{i,j,t}\left(\frac{\hat{\epsilon}_{i,j,t}}{\sum_{i,j,t}\hat{\epsilon}^2_{i,j,t}}\right), \tag{4}$$

with $\hat{\epsilon}_{i,j,t}$ representing the residual of regressing the treatment indicator in the effects fixed in the household, year, and quality. The effect of treatment is therefore a weighted sum of the outcome variable, where the weights are the residualized treatment weights. Jakiela (2021) indicates that bias arises when treated units have negative treatment weights and when treatment effects are heterogeneous. To identify such biases, we examine whether treated units have negative weights and then test for homogeneity of treatment effects.

In Appendix I.1, we provide a detailed analysis of the robustness of our findings, following the procedures outlined by Jakiela (2021) to show the weights for treated and untreated units. It is noted that only 15% of the treated units have negative weights. For comparison, Jakiela (2021) found that about 25% of the treated units had negative weights, yet the treatment effect remained robust after removing those observations. Since our estimate of the Average Treatment Effect (ATE) is a weighted sum of outcomes, these small negative weights are unlikely to cause bias. As an additional robustness check, we reestimated the model, this time excluding treated units with negative weights. Results in Appendix I.1 show that the effect of the platform launch on substitution remains significant.

4.5.2 Interaction Weighted Estimator

To avoid the potential for bias inherent in TWFE estimators, we have also implemented the interaction weighted (IW) fixed effects estimator, as proposed by Sun and Abraham (2021) and Callaway and Sant'Anna (2021). The IW estimator is robust to heterogeneous treatment effects in models with staggered treatment and can be used even without a never-treated group. According to the methodology proposed by Sun and Abraham (2021), our sample was divided into different cohorts based on the year that each household gained access to the platform. In our study, these results were obtained in three distinct cohorts (2018, 2019, and 2020), as well as a cohort that had not received treatment.



Notes: The above figure illustrates the impact of the GEP on online and offline tea sales for different cohorts. The blue dots indicate the effect on offline sales, while the red dots show the effect on online sales. The horizontal axis represents the periods since the treatment, and the vertical axis represents the estimated ATT with 95% confidence intervals. Although the figure clearly demonstrates that the trends of the groups were parallel before the intervention (parallel pre-trends), we also provide additional checks in Appendix H to further validate the assumption of parallel trends.

Figure 3. Average Treatment Effects by Cohort

In Appendix I.2, we provide a detailed analysis and results from the implementation of the IW estimator. As shown in Table I.2, our IW estimates confirm our initial findings on the impact of the GEP on tea sales. Converting our estimates to the effects on online and offline sales, we find that the GEP resulted in an average 14.444% decrease in offline sales and a 12.524% increase in online sales. Figure 3 shows the estimated effects across cohorts. We observe a consistent effect, indicating a significant positive impact on online sales and a negative impact on offline sales after the platform's implementation.

5. Heterogeneous Treatment Effects

Our heterogeneity analysis focuses on two dimensions: (1) differences between households with low vs. high production levels, and (2) differences across product quality tiers.

5.1 Effects Across Different Output Levels

We hypothesized that the impact of the GEP access could differ by farm size (output level). Specifically, the impact of the platform could differ for farmers with high and low output levels. To address this, we divided the households into three production-level groups based on the distribution of annual output in the sample: low output (0–405 kg/year), medium output (406–870 kg), and high output (870 kg). We perform subsample regressions for each category to assess two outcomes: the overall change in sales volume and the shift in sales from offline to online after the GEP launch.

The findings are presented in Table 3. It is indicated that the largest producers moved a greater share of their sales online than medium or small farmers. Additionally, we find that the introduction of GEP did not significantly change the total volume of tea sales. However, it led to a notable shift in sales from offline to online, particularly for the largest tea farmers. Structural factors (e.g., limited garden acreage and a finite number of ancient tea trees) limit how much tea output can grow. Such restrictions, particularly the lengthy maturation time required for high-quality tea leaves, hinder any rapid expansion in production. Changes in ownership among households were minimal, further indicating that total output remained stable before and after the intervention. Because total sales remain unchanged, the shift from offline to online appears to be a profit-maximizing reallocation by farmers to sales channels with higher margins or lower transaction costs.

Table 3. Heterogeneous Effects of the GEP Access on Sales by Quantity

Dependent Variable:	Log(sales): $q_{i,j,t}$							
	All Households		**0-405 Kg**		**406-870 Kg**		**871+ Kg**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Platform Access	-0.008	-0.169***	-0.006	-0.093***	-0.022	-0.182***	0.006	-0.230***
	(0.008)	(0.035)	(0.007)	(0.010)	(0.017)	(0.029)	(0.013)	(0.060)
Platform Access × Online Sales		0.323***		0.173***		0.319***		0.447***
		(0.078)		(0.019)		(0.056)		(0.140)
Online Sales	-0.392***	-0.482***	-0.125***	-0.175***	-0.393***	-0.477***	-0.689***	-0.816***
	(0.027)	(0.038)	(0.020)	(0.012)	(0.024)	(0.020)	(0.012)	(0.027)
Zero Output	-5.438***	-5.429***	-4.887***	-4.883***	-5.412***	-5.404***	-5.757***	-5.737***
	(0.069)	(0.065)	(0.058)	(0.056)	(0.031)	(0.029)	(0.041)	(0.032)
Observations	29,490	29,490	9,900	9,900	9,780	9,780	9,810	9,810
Household FE	YES	YES	YES	YES	YES	YES	YES	YES
Quality FE	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES
R^2	0.965	0.965	0.970	0.970	0.971	0.972	0.964	0.965

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. Significance levels are denoted as follows: ∗p < 0.10, ∗∗p < 0.05, ∗∗∗p < 0.01.

5.2 Effects Across Different Product Qualities

We also examine whether the impact of the GEP access varies by grade of tea quality. We classify tea into three quality tiers (premium, high, and regular) and run separate regressions for each in Table 4. Consistent with our previous findings, the GEP access again shows no significant impact on total sales volume across quality grades. Further analysis reveals that online sales of regular tea increased by 11.963%, high-quality tea by 15.604%, and premium-quality tea by 21.653%. Although the premium-quality tea segment shows the largest percentage increase in online sales, this is primarily due to its significantly smaller online sales volume compared to regular tea. In contrast, although online sales of regular tea increased by just under 12%, they experienced the largest volume increase (in kilograms per year) among online sales.

Table 4. Heterogeneous Effects of the GEP Access on Sales by Quality

Dependent Variable:	Log(sales): $q_{i,j,t}$							
	All Qualities		**Regular**		**High**		**Premium**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Platform Access	-0.008	-0.169***	-0.007	-0.127*	-0.010	-0.166***	-0.011	-0.219**
	(0.008)	(0.035)	(0.009)	(0.051)	(0.008)	(0.028)	(0.013)	(0.077)
Platform Access × Online Sales		0.323***		0.240**		0.311***		0.415*
		(0.078)		(0.092)		(0.049)		(0.175)
Online Sales	-0.392***	-0.482***	-0.398***	-0.469***	-0.139***	-0.407***	-0.461***	-0.579***
	(0.027)	(0.038)	(0.031)	(0.047)	(0.056)	(0.063)	(0.079)	(0.128)
Zero Output	-5.438***	-5.429***	-6.003***	-5.972***	-4.977***	-4.945***	-5.003***	-4.962***
	(0.069)	(0.065)	(0.047)	(0.040)	(0.142)	(0.142)	(0.078)	(0.057)
Observations	29,490	29,490	9,830	9,830	9,830	9,830	9,830	9,830
Household FE	YES	YES	YES	YES	YES	YES	YES	YES
Quality FE	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES
R^2	0.965	0.965	0.976	0.976	0.977	0.977	0.972	0.973

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. Significance levels are denoted as follows: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

6. Exploring Potential Mechanisms

We now investigate the mechanisms through which the GEP affects farmers' sales, focusing on the extensive margin (whether more farmers sell online) and the intensive margin (how the online presence of farmers expands). On the extensive margin, we examine the increase in the number of farmers selling online following the launch of the GEP. This increase is likely driven by the introduction of the platform and its accompanying public services (e.g., training and regional branding support).

6.1 Extensive Margin: Effects on Online Mode Adoption

We first look at the extensive margin: Does GEP access induce farmers who weren't selling online to start doing so? To assess the impact, we estimate three different adoption outcomes. First, we define an overall online adoption indicator ($Adopt_{i,t}$), which equals one if farmer i sells any tea online in year t, and zero otherwise. Column (1) of Table 5 shows that farmers who had no prior experience selling tea online before the introduction of the GEP are significantly more likely to start selling tea online after gaining access, with an estimated effect of 22.774% (significant at the 1% level).

In Columns (2) and (3), we further refine our analysis by examining adoption decisions using alternative metrics. In Column (2), the dependent variable ($Adopt_{i,j,t}$) equals one if farmer i sells tea of quality j online during year t. The results show a statistically significant increase of 7.551% (significant at the level 1%) in the probability that farmers adopt online sales for each tea quality category. Column (3) uses an alternative definition by estimating the effect of the platform on the timing of online adoption. Here, we use the dependent variable ($Adopt_{i,t}^{f}$), which is equal to one only if year t is the first year that the farmer sells tea online. The results indicate that farmers without previous online sales experience show a significant increase of 21.778% (significant at the 1% level) in the probability of making their first online sale after gaining access to the platform. These estimates suggest that the policy lowered entry barriers for households that had been offline, either by closing knowledge gaps about online sales or by helping them meet the onboarding requirements of commercial platforms.

Table 5. Effect of the GEP Access on Adoption of Online Sales

Dependent Variable:	Overall Online Adoption ($Adopt_{i,t}$)	Online Adoption by Quality ($Adopt_{i,j,t}$)	First-time Online Adoption ($Adopt_{i,t}^{f}$)
	(1)	(2)	(3)
Platform Access	0.228***	0.076***	0.218***
	(0.049)	(0.017)	(0.030)
Observations	745	745	745
Household FE	YES	YES	YES
Year FE	YES	YES	YES
R^2	0.548	0.205	0.357

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. The differences in observations across columns are due to conditioning on farmers who had no online sales experience before the introduction of the GEP. Significance levels are denoted as follows: ∗p < 0.10, ∗∗p < 0.05, ∗∗∗p < 0.01.

6.2 Intensive Margin: Effects on Number of Online Channels

Next, we consider an intensive-margin mechanism: Does the GEP help farmers expand their online presence across multiple channels? Specifically, we monitor the number of online channels each household utilizes. Before the policy, most online activity took place on social media, as large commercial platforms (e.g., Taobao or JD) often required deposits, fees, and higher operational capacity. In contrast, social media has lower entry requirements (see Appendix B for a detailed comparison). In line with a mechanism that reduces fixed and capability barriers through training, cooperative processing, and the presence of a GEP, we anticipate diversification into multiple online channels once access is granted.

Table 6. Effects of GEP Access on Online Channel Adoption

Dependent Variable:	Number of Channels	Adopt Social Media		Adopt Platform	
	Offline Only	**Offline Only**	**Excluding Social Media**	**Offline Only**	**Excluding Platform**
	(1)	(2)	(3)	(4)	(5)
Platform Access	0.228***	0.232***	0.234***	-0.004	0.116*

	(0.057)	(0.050)	(0.051)	(0.012)	(0.053)
Observations	745	745	760	745	4.095
Household FE	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES
R^2	0.562	0.548	0.547	0.351	0.437

Notes: Standard errors are indicated in parentheses. Standard errors are clustered at the area level. "Excluding Social Media" includes farmers with no online sales and those who sold only on the platform before the policy. "Excluding Platform" includes farmers with no online sales and those who sold only on social media before the policy. Significance levels are denoted as follows: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

The estimates in Table 6 show two patterns. First, among households without online sales before the policy, access to the platform increases the probability of adopting one additional online channel by 22.780%. Columns (2) to (5) further explain which channels are adopted. Restricting to households with specific pretreatment channel restrictions, access to the GEP increases the in-sample adoption rate of social media by 23.407% and the in-sample adoption rate of online platforms by 11.625%. Among households with no prior online sales, program access is associated with a higher likelihood of adopting social media channels, with no detectable change in the adoption of online platforms.

The results from Columns (2) to (5) convey an insight: When households first move online, they always prioritize the social media channel, even when the GEP is available. This is evident from the much larger coefficients for social media adoption relative to platform adoption. In the subsample estimates, the coefficient increase is 0.234 for adopting social media versus 0.116 for adopting the platform channel. Overall, the evidence shows that the estimated impact of gaining access to the GEP comes mainly from cross-channel spillovers of the bundled public service package: training, cooperative processing, packaging, and regional branding, rather than from the immediate, direct use of the GEP storefront alone. Therefore, even if the storefront were not maintained, farmers would continue to benefit from these public services and sustain profitable online sales through existing private channels. In addition, the adoption sequence aligns with the capacity and credibility constraints. Farmers typically begin by selling their products through informal, low-cost social media platforms; then, processing and branding services reduce participation costs and enhance credibility, enabling a gradual expansion into more formal marketplaces, including platform storefronts. This interpretation aligns with evidence that the reduction of participation barriers reshapes the structure of the market and entry in LMICs (Goyal, 2010; Bergquist and Dinerstein, 2020), and with the findings that platform assurances and third-party certification mitigate adverse selection in online markets (Jin and Kato, 2006; Saeedi, 2019).

6.3 Intensive Margin: Effects on Product Diversity

We also analyze product diversity on the intensive margin: Do farmers sell a wider range of tea qualities online after gaining access to the GEP? Since the market for regular tea tends to have thin margins in traditional online commercial markets, we hypothesize that the government's package of services - cooperative processing at minimal cost, regional public branding, training, and a commission-free sales channel - makes it cost-effective to list lowerpriced varieties online, while also allowing entry into higher-end segments for households previously absent from them. In Appendix J, we present some model-free evidence that corroborates this hypothesis.

Table 7 summarizes these two aspects of product diversity. First, we examine the number of varieties sold online and show that GEP access increased the overall variety of online tea products. Second, our results on adoption by quality show a clear reallocation along both ends of the spectrum. At the lower end, households that did not sell regular tea online before the policy are 5.759% more likely to do so after gaining access, even though the effect is nil for the previously offline group taken as a whole. At the upper end, households with no online sales and those without prior high/premium listings are 22.355% and 25.799% more likely, respectively, to sell high/premium tea online after gaining access.

Table 7. Effects of the GEP Access on Online Varieties

Dependent Variable:	Online Varieties	Online Regular		Online High/Premium	
	Offline Only	**Offline Only**	**Excluding Regular**	**Offline Only**	**Excluding High/Premium**
	(1)	(2)	(3)	(4)	(5)
Platform Access	0.228***	0.232***	0.234***	-0.004	0.116*
	(0.057)	(0.050)	(0.051)	(0.012)	(0.053)
Observations	745	745	760	745	4.095
Household FE	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES
R^2	0.562	0.548	0.547	0.351	0.437

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. “Excluding Regular” includes farmers who did not sell any regular-quality tea online before the policy. “Excluding High/Premium” includes farmers who did not sell any high-quality or premium-quality tea online before the policy. Significance levels are denoted as follows: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

These results show that program access broadens the online selection across both high-end and low-end products. We also observe a pattern in product adoption: Among households without prior online activity, the first products listed after gaining access to the GEP are mostly high- and premium-quality tea, while among households already selling online, access is followed by additional listings of regular tea. This pattern aligns with our expectations: Higher-grade teas have higher markups, making them the first natural products to go online. In contrast, regular tea usually has smaller online-offline price differences and is not prioritized without additional support. Since the GEP reduces online costs for selling regular tea through cooperative processing and compact packaging under regional public branding, it can help farmers list some of their regular output online and earn additional income.[5]

These intensive-margin responses match two strands of existing evidence. First, by lowering participation and operating costs for small producers, the public digital infrastructure can broaden market access and reallocate where transactions occur (Fernandes et al., 2019; Couture

[5] In Appendix K, we present additional evidence supporting this mechanism. When we restrict the sample to farmers who were already selling online before GEP access, we find that GEP access is associated with increased online sales across nearly all channel types. The exception is the subgroup already qualified to sell on commercial marketplaces before the program: For these farmers, GEP access has a non-significant incremental effect. They typically operate on a larger scale, meet qualifications for commercial platforms, and are almost universally active on social media; therefore, the GEP storefront does not significantly increase margins or profits. These findings collectively demonstrate that the GEP offers a cost-effective alternative channel, enabling farmers who previously could not profitably sell lower-grade teas on commercial platforms to market these products online.

et al., 2021). Second, reputation and certification alleviate quality uncertainty in online settings (Jin and Kato, 2006; Saeedi, 2019). In our context, standardized processing and regional public branding play a role analogous to third-party certification, reducing credibility costs for regular tea and supporting the listing of higher-end products by previously excluded households. Quality-side adjustments in response to access to distant buyers are also consistent with experimental evidence on product upgrade under export-market access (Atkin et al., 2017). The documented expansion of variety is consistent with our aggregate results. With the short-term inelasticity of the tea supply (fixed plot sizes and age requirements for tea trees), digital integration changes what is sold online rather than expanding total output (Foster and Rosenzweig, 2004). The public e-commerce service thus functions less as an online platform and more as an enabling infrastructure that lowers costs for a broader set of product varieties.

6.4 Mediating Role of Online Channels and Product Variety

To determine whether the observed margins are the primary channels driving the substitution pattern, Table 8 presents mediation regressions with the dependent variable defined as log of household–year level online sales aggregated across qualities. Column (1) repeats the baseline specification used in Table 2 for the aggregated online sales. Columns (2) and (3) add the number of online channels used by the household and the number of product varieties listed online separately. Column (4) jointly adds the number of online channels and product varieties together.

Table 8. Mediating Role of Online Channels and Product Variety

Dependent Variable:	Total Online Sales			
	(1)	(2)	(3)	(4)
Platform Access	0.081*	0.032	0.034	0.020
	(0.035)	(0.055)	(0.032)	(0.038)
Number of Channels		2.357***		1.217***
		(0.086)		(0.056)
Number of Varieties			2.579***	1.971***
			(0.062)	(0.036)
Zero Output	-2.387***	-0.953***	-1.031***	-0.610***
	(0.191)	(0.105)	(0.102)	(0.064)
Observations	4,915	4,915	4,915	4,915
Household FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
R^2	0.777	0.884	0.926	0.946

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. Significance levels are denoted as follows: $*$ p < 0.10, $**$p < 0.05, $***$p < 0.01.

Our results align with a mediated substitution mechanism. In Column (1), access to the GEP significantly boosts overall online sales.6 When we add the two potential mediators, each one enters positively and is highly significant on its own; both remain positive and statistically significant when included together, indicating they are not strongly collinear and each explains different variation. As we add these mediators, the coefficient on GEP access shrinks toward zero and becomes statistically insignificant by Column (4). At the same time, model fit improves markedly once the number of online channels and the number of product varieties are included.

This pattern supports the idea that GEP access primarily increases online sales by expanding the number of online channels and broadening the range of product offerings available online. In Appendix M, we present parallel analyses using aggregate offline sales as the outcome. We also control for the number of shipping companies to ensure that the increase in online sales is not simply due to simultaneous growth in local shipping capacity. We find that including shipping companies as a control does not change our results in this section. We show that the same mediators, more online channels, and an increased variety of online products are also associated with greater declines in offline sales, and including them significantly reduces the coefficient related to access to the GEP.

7. Conclusion

This chapter examines the impact of a government-initiated e-commerce program on the sales of rural producers, considering both offline and online channels. Using a staggered county rollout in Pu'er, China, and a near-census household panel of 983 tea producers from 2016 to 2020, we employ TWFE models to distinguish between online and offline responses at the household-year-quality level. We show clear evidence that access to the GEP reallocates sales across channels. Online sales increase by approximately 16.649%, while offline sales decrease by approximately 15.549% after gaining access, resulting in no overall change in total sales in the short to medium term. This pattern remains consistent when using alternative estimators that account for staggered treatment.

Mechanisms operate on both margins. On the extensive margin, program access increases the likelihood that previously offline households start selling online. On the intensive margin, sellers broaden their online footprint across channels, typically entering via social media formats and then adding online platforms, and expand the range of tea qualities they offer. Adoption occurs at both ends of the quality spectrum: households that had not listed regular tea begin doing so, and households without prior premium listings add higher-grade products. Mediation tests reveal that the policy effect is primarily transmitted through two pathways: increased online sales channels and expanded online product variety.

Two policy implications follow. First, where supply is inelastic in the short run, public digital services and government-initiated platforms can reallocate transactions from offline to online channels, thereby widening participation on the extensive margin. The mechanism operates by lowering fixed and capability costs through training, shared processing and packaging, standardized quality control, and public regional brands. In LMICs, these features are particularly important for smallholders who face thin margins, limited buyer recognition, and restricted access to professional e-commerce operations. Second, the public sector should complement rather than replace private marketplaces. Useful complements include community processing facilities, seller training, and basic digital infrastructure for collection and delivery. These packages may offer lower-margin varieties online and encourage multichannel strategies for households and cooperatives. In short, digital public services shift what and where is sold initially, while long-term gains in scale and income can occur when complementary investments relax the current supply-side constraints.

While the GEP does not appear to increase total tea sales in the short to medium run, this does not necessarily imply that farmers experience no welfare gains. The platform and its bundled public services may still improve farmer welfare by lowering transaction and operating costs, reducing commissions, expanding market access, and improving credibility through cooperative processing, packaging, and regional branding. In this sense, the main effect of the GEP may be to reallocate sales across channels while also changing the cost and quality of market participation. At the same time, the present analysis does not directly observe net household income, profits, or consumption, so the chapter cannot make a strong causal claim about the overall magnitude of welfare gains. A fuller welfare evaluation would require additional data on prices received, production costs, household income, and longer-run adjustments in production capacity.

Chapter 2: Referral Networks and Renewal Rates: An Empirical Study from a Creator Platform

The success of digital products increasingly hinges on both customer acquisition and retention. According to a Parks Associates study of the top 10 U.S. subscription over-the-top video services in 2022 (Sorensen, 2022), the most successful streaming platforms tend to retain customers for longer periods of time. Users of Netflix and Amazon Prime, for example, often maintain their subscriptions for an average of four to five years. Early literature suggests that retaining customers generally costs less than acquiring new ones and provides additional benefits by improving overall lifetime value (Fornell and Wernerfelt, 1987). However, the digital era has shifted user behaviors. With the ease of signing up and canceling services, businesses often need to exert more effort to retain their customers. According to Apptentive's 2021 Mobile App Engagement Report (Alchemer, 2021), a study of 1,000 apps with over one billion app installs, retention has become a primary business focus for mobile app developers. The report found that across all apps, only 35% of consumers who used the app in January were still active users in December. With this in mind, business practitioners are looking for digital tools to improve user retention for digital products. Among these tools, referral programs have received significant attention from business practitioners and scholars over the past decade as an effective means to improve user retention. For example, a Forbes report in 2015 (Newman, 2015) cited market data indicating that customers acquired through referrals have a 37% higher retention rate. Similarly, Armelini et al. (2015) found in their study that the risk of customer churn is also reduced with referrals. Despite the general consensus on the effectiveness of referral programs in improving digital product retention, there is still a need to explore the complex relationship between referral programs and user retention rates in greater depth. Digital platform managers are grappling with key questions: 1) Are referral programs an effective tool for increasing digital product subscription renewal rates? 2) To what extent are users' renewal decisions influenced by their referral network peers? 3) From a commercial perspective, which referral network structures, such as one-to-one relationships or influencer-centric networks, offer greater benefits to digital platform managers and business strategists? And 4) Within referral networks, who should be the primary target of loyalty campaigns: the referrers or the referees? Understanding and answering these questions is critical to realizing the full potential of referral programs to drive customer retention, and forms the cornerstone of our study.

To address these questions, we need an appropriate empirical framework to isolate the effects of being referred (as a referee) and referring others (as a referrer) within the referral network. Furthermore, to accurately quantify the complex changes induced by network shifts and platform policies on individual agents in the referral network and their ultimate impact on aggregate-level customer retention, we need to simulate the decision-making process of users within the referral network at the time of renewal through a structural model. This approach also allows us to conduct meaningful counterfactual analyses of business strategies. Beyond the data and modeling challenges, we also face significant empirical challenges in identifying effects. These challenges center on mitigating endogeneity concerns in both pricing and network linkages. Such concerns are particularly complicated in the context of referral networks, where the interaction of various unobserved factors can significantly affect the interpretation of model results. Therefore, careful consideration and resolution of these issues is fundamental to the robustness of our study.

In this study, we use data collected directly from the KnowledgePlanet platform to examine the impact of referral relationships on users' decisions to renew their digital content subscriptions. KnowledgePlanet is a prominent content delivery platform in China, where users pay annual fees to subscribe to various communities. These communities, each managed by an owner, offer different content and have different pricing policies, with subscription fees ranging from 50 RMB to 5,000 RMB.[6] One notable aspect of KnowledgePlanet is the implementation of a special referral program. Under this program, users who refer others to join a community share the subscription fee with the community owner. In addition, those who join as referees through such referrals receive a discounted annual fee upon payment. This referral program is designed to benefit both the referrer and the referees, and as such, the platform effectively monitors and records all referral activity. Capturing the entire network is essential to accurately assess how different influences propagate through it, as partial views of the network can lead to biased estimates due to missing observations, a point highlighted in studies such as Demir et al. (2022). Another advantage of our data is that each community owner has the freedom to set and change renewal prices in their community independently. This autonomy allows us to observe different renewal prices for the same community at different times, enriching the data with different price variations. This diversity in pricing allows us to simultaneously identify the influence of both price changes and referral relationships on users' renewal decisions, providing a comprehensive understanding of the dynamics at play in these digital communities.

Our dataset contains detailed records on over 200,000 users in 173 randomly selected communities. It includes detailed records of user payments, time-varying characteristics of each community, and most importantly, each user's referral relationships within those communities. Using the referral information, we compute the referral network for each community. Taking advantage of the rich observational data and the unique empirical setting, we develop linear regression models and corresponding structural models. These models allow us to examine the various factors that influence users' renewal decisions, tailored to specific model specifications. Specifically, we focus on the effects of price, whether a user was referred, the referrer's decision to renew, and the number of users a member referred to the community in the period prior to their own renewal. The first variable captures the effect of price variation, while the latter variables capture the effects induced by the dynamics of the referral network. In our structural model, we also allow users to rationally predict the probability that their referrals will renew in the following period. By incorporating this into our estimation process, we further capture how dynamic price changes can affect the focal user's expectations and thus the renewal decision.

The empirical study reveals several key findings regarding the dynamics of price changes, referral relationships, and network structures in influencing user renewal decisions on digital platforms. First, we calculate the price elasticity associated with users renewing their digital memberships. At the same time, our analysis reveals a reciprocal positive influence on the focal user's renewal decision, suggesting that social enrichment is likely to play an important role in improving customer retention. We find that users are more likely to renew their subscriptions

[6] The median per capita disposable income in China was around 27,540 RMB (approximately 4,262 USD) in 2020, and 1 RMB is equivalent to about 0.15 USD.

when two key conditions are met: 1) their referrers have already renewed, and 2) they expect their referrals to renew as well. In addition, our results suggest that referred customers are not necessarily a better fit for the product: They are more likely to leave their communities. This finding differs from previous studies that report a negative correlation between referrals and churn rates (Beaman and Magruder, 2012; Brown et al., 2016; Kornish and Li, 2010; Pallais and Sands, 2016). Our results are consistent with the findings Belo and Li (2022), which highlight that while social referral programs may increase referrals, they may also reduce user engagement and thus impact the overall value of the platform. Our estimate of these effects remains robust and unchanged even after conducting thorough endogeneity checks on both pricing and referral network dynamics.

Second, our counterfactual analysis examines three pricing strategies: discounts for all users, discounts for referrers only, and discounts for referrers only. By applying uniform discounts to users, our counterfactual results show that the network spillover benefit of discounting is concave, reaching its maximum impact at discounts between 5% and 10%, which subsequently increases the renewal rate by an additional 4%. Such a pattern reveals a complex interplay between price and network effects. For smaller discounts, renewal decisions are primarily influenced by the user's immediate network contacts. However, as discounts become larger, a larger number of users are influenced, amplifying the network effect. However, when price reductions are significant, the primary driver of renewal decisions shifts to the lower price itself, reducing the relative influence of network peers. In this context, we find that offering targeted discounts to referrers, rather than uniform discounts to all subscribers or exclusively to referrers, is a more effective strategy. Given the estimated level of price elasticity, we show that focusing renewal discounts on referrers is a more efficient strategy for increasing renewal rates while reducing the revenue loss associated with price reductions.

Third, we further investigate which network structure is most effective in increasing a platform's average renewal rate. We focus on two key variables: connectivity and centrality. The results of our counterfactual analysis show that networks with high connectivity and low centrality are more successful at maintaining high customer loyalty and retention. This finding differs from the traditional emphasis on targeting super-influencers in business operations (Hinz et al., 2011). Instead, our findings argue for a strategy that encourages each user to leverage his or her influence sequentially in a chain-like fashion, which significantly improves user retention. In networks where referrers and referents influence each other, a more centralized structure may reduce the hierarchical breadth of this influence, thereby limiting the frequency and magnitude of positive effects as they spread through the network. This scenario could potentially reduce the overall effectiveness of the snowball effect (Hada et al., 2014). Further empirical evidence suggests that community owners tend to prefer lower levels of network centrality, especially when managing large user bases.

Our study makes several important contributions to the literature. First, it offers new perspectives on promotional strategies and referral programs, areas rich in both theoretical and empirical studies (Biyalogorsky et al., 2001; Garnefeld et al., 2013; Fernández-Loría et al., 2023; Hinz et al., 2011). Our study is distinguished by proposing a structural model that effectively integrates these two domains. This model examines the balance between the economic value derived from referral networks and the costs of offering discount incentives, providing a comprehensive view that links promotional strategies with the design of referral systems. This approach provides detailed insights into the cost-benefit aspects of referral program design, thereby deepening understanding in these areas and laying the groundwork for future marketing and information systems research. Second, our study adds to the literature on how referral relationships influence consumers' decisions to subscribe to digital products. It considers two potential mechanisms: better matching and social enrichment. The better matching mechanism, as suggested by previous studies (Beaman and Magruder, 2012; Kornish and Li, 2010), suggests that referred customers are more likely to find a product or service that meets their needs. In contrast, the social enrichment mechanism, as suggested in works such as Schmitt et al. (2011) and Van den Bulte et al. (2018), highlights the benefits of a more integrated and closed social structure for both referrers and referees. Our study suggests that the influence of network peers on average renewal rates is likely to be more complex and interactive than previously recognized. In our dataset, the low cost of referrals on digital platforms may lead to higher churn rates for referred customers who are not well suited to the product. Users are more inclined to renew their subscriptions only when the number of referrers and referees who decide to renew increases. Finally, our study methodologically advances the study of network interactions in marketing and information systems. Departing from the network studies (Bhattacharya et al., 2019; Hu et al., 2019; Wei, 2020), our structural approach quantifies the effects of different pricing policies by incorporating users' expectations of their network peers' decisions. This approach highlights the importance of network interactions for effective user relationship management, as highlighted by Ben Rhouma and Zaccour (2018). In addition, our empirical findings provide strategic insights for business practitioners seeking to influence the development of referral networks. By limiting the number of referees a user can refer, digital platform owners with larger customer bases are encouraged to form less centralized networks. Strategies that optimize the potential structure of referral networks would increase the long-term profitability of the platform by increasing renewal rates.

1. Related Literature

1.1 Referral Networks and Customer Retention

The exploration of retention and loyalty programs, along with price promotions, has been a central theme in research of marketing and information systems, with scholars aiming to understand and enhance customer retention and loyalty (Bolton et al., 2000; Leenheer et al., 2007; Hristakeva and Mortimer, 2023). In parallel, the significance of social networks in boosting customer engagement and facilitating product adoption has been increasingly recognized, as evidenced in studies by Hinz et al. (2014) and Shriver et al. (2013). This growing understanding 7 has led to the integration of social networks into retention and loyalty strategies, reflecting the evolution of referral programs alongside the digital revolution. Traditional word-of-mouth tactics have transformed into more advanced, digital strategies, particularly within the realm of social media, significantly enhancing the scope and impact of referral programs (Galbreth et al., 2012; Sun et al., 2020). Empirical studies on referral networks has bifurcated into two primary areas of focus. One examines the benefits referral networks offer to firms, such as pinpointing influential agents (Roelens et al., 2016) and augmenting customer value (Schmitt et al., 2011; Van den Bulte et al., 2018). The other focuses on the structural design of referral programs, exploring elements like social contagion effects (Aral and Walker, 2011), boosting referral probabilities (Jung et al., 2020; Jung et al., 2021; Ryu and Feick, 2007), and the crafting of effective referral messages (Fernández-Loría et al., 2023).

Our study bridges these two streams of literature by proposing an empirical framework that balances potential revenue loss with customer retention and loyalty. Unlike most prior literature that focuses on exploring the effectiveness of referral programs primarily based on customer acquisition (Biyalogorsky et al., 2001; Fernández-Loría et al., 2023; Garnefeld et al., 2013), our study integrates the realms of promotion strategies and referral system design within the framework of customer retention. We explore how the economic benefits derived from referral networks can be effectively balanced against the costs of providing renewal discount incentives. Our empirical findings also resonate with theoretical predictions from scholars such as Kamada and Öry (2020) and Lobel et al. (2016). Kamada and Öry (2020) suggest that overly strong referral incentives may backfire, diminishing the expected benefits for referees and, consequently, impacting firms' revenue adversely. Lobel et al. (2016) propose that the optimal referral payment should be non-monotonic in relation to the number of successful referrals, a concept our study further elucidates. Through this integration, our study introduces a structural modeling approach that enables digital platforms to not only evaluate how pricing and referral policies influence retention rates through referral networks but also to optimize their referral programs through counterfactual analysis to maximize long-term profitability.

1.2 Network Effect Identification

Network effects play an essential role in the digital landscape (Boudreau et al., 2022; Goldfarb and Tucker, 2019; Kapoor et al., 2018). While network analysis is recognized as critical (Reingen and Kernan, 1986), much existing studies have predominantly focused on the effects of direct peers within networks (Bailey et al., 2022; Chu and Manchanda, 2016). Concurrently, other studies have concentrated on how network-related statistical variables influence business outcomes. For example, Muller and Peres (2019) highlighted the importance of network aspects such as closeness and centrality, noting that networks with high cohesion, connectedness, and conciseness are more likely to promote growth. However, the influence of network structure on the propagation of effects within the network is often overlooked due to several challenges. First, the dynamic evolution of complete network structures is difficult to observe, leading scholars to rely more on psychographic analyses (Campbell et al., 2014; Hinz et al., 2011). Second, the decision-making process within networks often involves complex interactions with network peers, which poses significant challenges for causal identification. For example, Hartmann et al. (2008) pointed out that endogenous group formation, correlated unobservables, and simultaneity are key confounding factors in this context. Finally, the computational demands associated with the analysis of large networks also pose significant challenges, as discussed by Mele (2022).

In this study, we address existing gaps by introducing a practical methodology for analyzing the influence of network structure on customer renewal decisions. Our empirical study uses an extensive user-level dataset that thoroughly records each user's referral relationships, allowing us to model their renewal decision process. This strategy allows for detailed tracking of decision propagation within the referral network, distinguishing between static variables (such as the mode of joining a community, whether by referral or not) and dynamic variables, specifically the renewal decisions of referrers and referees. Unlike other structural approaches in the context of networks (Bhattacharya et al., 2019; Hu et al., 2019; Wei, 2020), our structural model allows for endogenous decisions within the referral networks, paralleling methods such as Ackerberg and Gowrisankaran (2006). A key feature of our study is the use of referral network data to identify the order in which users join and their corresponding decision timelines. The sequential nature of the decision-making process in referral networks greatly enhances the precision of our estimates while significantly reducing the complexity associated with the computational processes. Our empirical approach not only strengthens the accuracy of our results, but also contributes to the broader discourse on the role of network structures in shaping consumer behavior in digital markets.

1.3 Product Matching and Social Enrichment

In the context of customer retention and loyalty, two primary mechanisms—better matching and social enrichment—emerge as focal points of interest. The concept of better matching has been a subject of study across various disciplines, including sociology, economics and information systems (Beaman and Magruder, 2012; Hildebrand et al., 2013; Pallais and Sands, 2016). In the literature, Van den Bulte et al. (2018) describe two forms of better matching: active and passive. Active matching involves a deliberate selection by the referees, who use their knowledge of both their peers and the firm's offerings. This process may involve costs associated with searching for and evaluating candidates, resulting in a more accurate fit between the referred individual and the product or service (Kornish and Li, 2010). Passive matching typically results from homophily, where individuals naturally gravitate toward those with similar characteristics. On the other hand, social enrichment, as explored in studies by Ansari et al. (2018), Schmitt et al. (2011) and Van den Bulte et al. (2018), emphasizes the strengthening of a customer's relationship with a firm through the intervention of a third party who is connected to both. This triadic relationship can also facilitate practical benefits, such as knowledge sharing and transfer (Bursztyn et al., 2014; Inkpen and Tsang, 2005). The interplay between better matching and social enrichment can be complex. When referral costs are high, both mechanisms tend to be aligned, as referees carefully select their referees to reduce the likelihood of mismatch. Hong et al. (2017) argue that both fairness and social distance influence users' referral behavior and their willingness to accept recommendations. Therefore, businesses should adopt differentiated strategies in designing referral reward mechanisms based on varying social relationships to enhance referral success rates and user engagement. Huang et al. 2022 use data from 1,684 referrers on a WeChat social network, tracking their sales performance and social interactions with referees over an 11-week period. The results show that sellers' effort and performance are jointly affected by social influence, competitive effects, and free-riding behavior. Belo and Li (2022) find that under strong monetary incentives for referrals, where reputational concerns are minimal, the matching mechanism offsets the benefits of social enrichment. In their study, when referrers are incentivized to increase referrals to gain access to a platform's freemium features, user engagement declines, ultimately diminishing the platform's overall value.

Our study provides empirical insights into how two mechanisms that may exist within referral networks, namely social enrichment and better matching, influence users' decisions to renew subscriptions to digital content products. We interpret them using different empirical patterns in the data. Evidence on whether referred users are more or less likely to renew speaks to the better-matching mechanism, while evidence on how a focal user responds to the renewal decisions of referrers and expected downstream referees speaks to the social-enrichment mechanism. We find evidence that referred users are not necessarily better matched to the product. However, these users often continue their subscriptions, influenced by the renewal decisions of their peers within the referral network. This finding illustrates the reciprocal nature of social enrichment, where the decisions made by referrers and their expected impact on those they refer are interdependent. Accordingly, these findings should be interpreted as evidence that is more consistent with social enrichment than with better matching in this empirical setting, rather than as a strict causal separation of the two mechanisms. Our study adopts a dual-perspective approach, which provides a novel perspective on the study of referral processes. This approach distinguishes our work from studies such as Hinz et al. (2014), which primarily examine the dynamics of peer pressure in advisor-advisee networks. As such, our study contributes to a broader understanding of the impact of referral network dynamics on client loyalty.

2. Institutional Background

Our study focuses on "KnowledgePlanet," a leading Chinese online platform for text-based knowledge and expertise sharing that was launched in 2015. By 2022, it has amassed nearly 50 million registered users, with daily active users of around 500,000. The platform is structured around "communities," each consisting of an owner, subscribers, and mostly text-based content. Similar to emerging content monetization platforms such as Patreon or BuyMeACoffee, KnowledgePlanet functions as a marketplace. It enables community owners, who are often content creators with a substantial follower base on other social platforms, to monetize their knowledge or expertise. These owners leverage their existing follower base to promote their communities, creating an initial user base that paves the way for revenue generation.

These communities are the core of content delivery. Each community is managed autonomously by its owner, who has the discretion to set subscription prices, offer discounts, and implement specific policies. Our study focuses primarily on paid communities, where users pay an annual subscription fee to gain full access to the community's content. Specifically, we look at communities where at least one user renews their subscription. This focus implies that the selected community owners have been successful in not only attracting, but also retaining their customers. Other subsections explore the platform strategies and tools used by community owners to attract and retain users, particularly as their subscriptions approach expiration. In particular, KnowledgePlanet enforces strict user registration protocols that link users to personally identifiable information to prevent the creation of multiple accounts. This system effectively prevents users from rejoining communities as new users.

The platform's referral program and subscription renewal system are instrumental in our empirical identification process and play an important role in understanding user behavior and community dynamics.

2.1 Referral Program

KnowledgePlanet uses a referral program as a key strategy for community owners to grow their user base. This program gives owners the discretion to set a referral incentive rate, ranging from 0% to 50%. This rate determines the financial rewards for both the referrer and the referee. An example of how the referral program works is illustrated in Table 9. For a community with an annual fee of $100, a 50% referral incentive, and a 20% renewal discount, the dynamics are as follows: An existing user (referrer) sends a referral link to a potential new user (referee). The referrer joins the community by paying the $100 annual subscription fee. Of this fee, the community owner receives 50% ($50). The remaining $50 is split between the referrer and the referee, with 70% ($35) going to the referrer and 30% ($15) to the referee, according to the ratio established by the platform.

Table 9. Pricing under $100 Initial Price, 50% Referral Reward, and "20% off" Renewal Discount

	Owner ($)	Referrer ($)	Referee ($)	New User without Referral ($)
Year 1 (with referral)	50	35	-85	
Year 1 (without referral)	100			-100
Year 2	80		-80	-80

Notes: In practice, community owners and users will have to pay an additional 30% commission fee to the platform for all transactions incurred, which we will not consider here.

A critical aspect of this program is that a user can only send referral links after joining a community. This design allows both the referrer and the referee to earn monetary rewards. The platform effectively monitors users who join communities via referral links, allowing for accurate mapping of the referral network within each community. It is important to note that the financial incentive for referring a new member is only for the initial subscription period. During the renewal period, neither the referrer nor the referee will receive additional financial benefits based on their referral status. This policy delineation is crucial for our study, as it helps distinguish the influence of the referral network from price effects when analyzing users' renewal decisions.

2.2 Renewal Notifications and Renewal Discount

KnowledgePlanet enables community owners to offer renewal discounts as a customer retention tool. Owners can set the discount rate at their discretion, and this offer is prominently displayed on the user interface during the renewal process. As a subscription approaches its annual expiration date, members receive timely renewal reminders. Failure to renew results in loss of access to community content. For clarity in our analysis, we consistently use the "they" to represent all users in our data set.

Figure 4 illustrates the renewal mechanism. As Table 9 shows, "new users" become "old users" one year after joining, at which point they face a renewal decision. For example, with an 80% renewal discount (effectively a 20% discount), they would pay $80 to renew. It is important to note that the benefits of the referral program only apply to the first year of membership; after the second year, referral status has no effect on renewal pricing.

The design of these renewal mechanisms supports the credibility of our study. First, renewal reminders are sent only 14 days before expiration, minimizing the likelihood of early renewals due to temporarily low prices. Second, the platform allows community owners to freely set the discount rate, resulting in price variation within the same community. These variations help identify price elasticity. Third, unlike platforms like Netflix where subscriptions are automatically renewed, KnowledgePlanet requires a conscious decision to renew. If a user does not respond to the renewal prompt, it is considered a decision not to renew, and there are no additional fees or costs associated with this decision. This feature ensures that users on our platform are renewing their subscriptions after careful consideration. Finally, the annual renewal cycle, similar to services such as Amazon Prime, reduces time dependency and encourages more rational decision making. Compared to monthly subscriptions (e.g., Netflix, Spotify), an annual fee is larger and typically requires more deliberate decision-making on the part of the user.

Notes: The figure illustrates the renewal process for a community's subscription service. Top block: When subscriptions are about to expire, users receive a "Service About to Expire" alert with two options-close the window or proceed with renewal. Bottom block: Users who choose to renew are taken to a page that displays the renewal price (e.g., "199 RMB/year"). The process concludes with the user clicking "Pay Now" to complete the transaction.

Figure 4. Illustration of the Renewal Process

3. Data and Regression Evidence

Our dataset is derived from the KnowledgePlanet platform, focusing on communities where at least one user has renewed their subscription. To the best of our knowledge, this proprietary dataset has not been used in prior academic studies. By 2020, the platform boasted nearly 2,000 active communities with over 2 million paying users. Communities are categorized by content topic, and the platform generates terabytes of data daily, which is stored on cloud servers. Given the size of the user base, structural model analysis of the entire sample is impractical. In addition, the data protection policy of the platform restricts our access to the comprehensive user data. Consequently, we engaged in extensive discussions with the platform's data team to obtain a dataset suitable for our study.

After a negotiation process, we are allowed to collect data from communities that focus on providing business and science content. Our selection criteria were twofold: First, the communities had to be established for more than 1.5 years to ensure that a significant number of users had made at least one renewal decision. Second, these communities had to have more than 50 users and recent content updates by the owner to show that the community was actively providing content.[7] Random sampling is used to overcome computational challenges and to gain meaningful insights from a representative sample. From over 600 communities that met our criteria, we randomly selected one-third for analysis in consultation with the platform. This process resulted in a final dataset of approximately 200,000 users from 173 randomly selected communities, culminating in a sample of over 260,000 observations. These users joined different communities on the platform between 2016 and 2020, with some renewing multiple times and others stopping payments after a year.

One empirical challenge is that renewal prices for users who choose not to renew their subscriptions are not observable. Each user may receive a different renewal price at each stage because the community owner adjusts the pricing dynamically over time based on content updates or other factors. To address this challenge, we conducted an in-depth analysis, which is detailed in Appendix N. We examined different price imputation windows (w = 3, 7, 14, and 30 days) and evaluated their impact on our empirical results. Our results indicate that a 14-day window (±14 days from the expiration date) strikes an optimal balance, as confirmed by regression models that include critical variables and fixed effects. This window choice maximizes data recovery while preserving the precision of our imputed prices, as indicated by the improved R-squared values and stable regression coefficients observed in our models. After documenting the variation in regression coefficients and matched data proportions across different window widths, we ultimately selected the 14-day window as our standard for subsequent data construction and empirical investigation. This decision reflects our commitment to balancing the precision of imputed prices with the comprehensiveness of our dataset to ensure the reliability and robustness of our empirical results.

The extent of multi-homing among users also poses a challenge to our analysis, as it suggests that users are considering multiple community memberships when making renewal decisions. This adds complexity to the interpretation of users' renewal behavior. Therefore, assessing the degree of multi-homing is crucial. To address this, we first examine the extent of multi-homing among users. As detailed in Appendix O, our investigation shows that only about 2.21% of users engage in multi-homing. Consequently, in our subsequent analysis, we abstract from the possibility of multi-homing and focus on the behavior of individual users within individual communities. Accordingly, the outcome in our analysis should be interpreted as renewal or non-renewal within a focal sampled community, rather than as a direct measure of whether a user remains active elsewhere on the platform.

[7] Contents in the platform spans eight distinct fields: Art, Economics, Education, Entertainment, Fashion, Health, Life, and Science. Among these, economics and science-focused communities are particularly prominent, drawing users who tend to exhibit cautious and rational payment behaviors.

3.1 Variables and Summary Statistics

Our sample primarily consists of economics-related communities, making up 91% of the total, while the remaining 9% are science-related. Summary statistics of key variables in our dataset are presented in Table 10. On average, a typical community in our study has 1,691 users and has been active for about three years. These communities are remarkably dynamic, with a total of over 4,900 articles posted over three years, averaging about 4.48 articles per day. In addition, community owners actively engage with their members, answering an average of 1.36 questions per day. The relatively high volume of articles and answers indicates that the community owners in our sample are diligently managing and interacting within their communities. In each community, each referrer recommends the community to an average of four referees. Although the numbers of referrers and referees are highly skewed, the analysis is based on 173 randomly selected communities, and the main results remain similar in specifications that include community fixed effects.

Table 10. Summary Statistics of Key Variables for Communities and Users

Statistics	**Mean**	**St.Dev.**	**Min**	**Max**
	Communities (N=173)			
Number of Users	1,690.66	5,191.79	51.00	59,021.00
Number of Articles	4,916.00	8,745.50	111.00	74,692.00
Number of Answers	1,497.00	3,935.39	0.00	26,399.00
Community Age	2.98	0.77	1.67	5.17
Network Statistics				
Number of Referrers	38.00	364.86	0.00	5,138.00
Number of Referees	141.80	1,352.54	0.00	17,794.00
	Observations (N=263,161)			
Joining Price (RMB)	334.70	450.18	0.00	5,888.00
Renewal Price (RMB)	353.25	496.04	25.00	5,888.00
Renewal Decision	0.54	0.50	0.00	1.00
Renewal Time	0.81	0.85	0.00	4.00
Smartphone Model	0.01	0.12	0.00	1.00
Referred	0.07	0.25	0.00	1.00
Number of Referees	0.08	4.22	0.00	1,309.00

Notes: Our statistics are based on observation-level data, which means that a single user may be counted multiple times in different observations. As a result, the actual number of user-level referrers and referees may be slightly lower than what we show above. Data compiled in June 2021 includes “Community Age” (duration since each community’s inception), “Number of Referrers” (community members who made successful referrals), and “Number of Referees” (those joining through referrals and completing payment). The observational exchange rate was US$0.15 to 1 RMB. The base renewal price was 50 RMB, with potential discounts down to 25 RMB. iPhone users are coded as 1; others are coded as 0. “Referred” is 1 for users joining through a referral link.

The average initial joining price for users in a community is 334 RMB, while the average renewal price is slightly higher at 353 RMB. This increase in the renewal price can be attributed to the fact that community owners generally raise prices over time, reflecting the development and growth of their communities. From an observation level perspective, referees account for only 8% of our dataset, or roughly one arbitrator for every 13 observations. The overall renewal rate in our sample is 54%, with an average renewal frequency of only 0.81 times per user. iPhone users represent a small fraction of the total user base, accounting for less than 1%.

3.2 Regression Model Specification

We begin by exploring the interactions between pricing, referral networks, and renewal decisions by employing a Linear Probability Model (LPM). For a given user i belonging to community m and facing a renewal decision at time t, the following LPM is considered:

$$d_{i,m,t} = \beta_0 + \underbrace{\beta_p \times \ln Price_{i,m,t}}_{\text{Price Effect}} + \underbrace{\beta_R \times R_{i,m,t} + \beta_{RD} \times RD_{i,m,t} + \beta_{NR} \times \ln NR_{i,m,t}}_{\text{Network Effect}} + u_{i,m,t}, \quad (5)$$

where $d_{i,m,t}$ represents a binary outcome variable, indicating whether user i opts for renewal within community m at time t; $price_{i,m,t}$ is the renewal price presented to the user at the decision-making juncture; $R_{i,m,t}$ indicates whether user i has a referrer or not at time t; $RD_{i,m,t}$ is a binary variable reflecting whether the referrer of user i renews during time t; and $NR_{i,m,t}$ signifies the count of referees who are due for a renewal decision in the subsequent year. This variable is intended to capture the focal user's downstream network exposure. It matters because users may derive value from continued interaction with the referees they have brought into the community, so a larger set of downstream referees approaching renewal may increase the focal user's own incentive to renew. $u_{i,m,t}$ is an unobserved error term. In the above equation, β_p captures the effect of price changes; β_{RD} captures the linear correlation between the renewal decision of user i and her referrer; β_{NR} captures the effect of having more referees in the same community on the focal user's renewal decision at t (measured in year-month). The baseline specification uses the number of referees because it is observable at the time of the focal user's renewal decision, whereas the realized renewal rate of referees is not yet observed at that point.

3.2.1 Additional Control Variables

Our regression model, shown in Equation 5, primarily examines the relationship between price, network variables, and user renewal decisions. However, to gain a more refined understanding of the effect of price and referral networks on user behavior, it is essential to consider additional variables present in our panel data. To this end, we introduce a set of additional control variables into our regression model to account for various underlying factors that may influence user decisions.

Fixed Effects to Address Time-Specific and Community-Specific Factors. User actions such as joining or renewing memberships may coincide with special promotional events on the platform. These events can significantly influence user behavior, particularly in referral networks where users may be more inclined to refer others during promotional periods. In addition, such events may also affect the pricing of subscriptions, thereby influencing renewal decisions. To control for these potential confounding factors, we introduce fixed effects at the year-month level that refer to the specific times when users join or renew their memberships. This adjustment allows us to isolate the effects of price and network variables from those of time-specific promotional influences, including those that may affect renewal behaviour through channels other than price. However, the available data do not separately identify community-specific non-price promotional activities beyond observed renewal pricing decisions.

The dynamics within each community can vary considerably. For example, in some communities, users may be more likely to refer others, and these users may also be more likely to renew their subscriptions. Failure to control for these community-level factors could exacerbate the endogeneity problems associated with network-related variables. In addition, certain communities may inherently have a more loyal user base. To account for these intrinsic community-level characteristics, we include community-level fixed effects in our regression analysis. These fixed effects help us control for unobserved community-specific characteristics that might influence user renewal behavior and referral activity. These fixed effects also help absorb systematic differences across user cohorts, although the current specification does not explicitly model subscription tenure or habit formation in renewal behavior.

User Characteristics. While a large set of fixed effects largely controls for unobservable factors that might bias our estimates, we still account for specific user characteristics that might influence renewal decisions. For example, some user – particularly those who joined through referrals – may have joined the community at a lower price during special promotional events in their first year. These users may have a different perspective on price and renewal decisions than those who joined at regular prices. To account for this, we include controls for initial joining price in our regression model. This helps to mitigate any impact this variable might have on our price elasticity estimates, especially considering that users might evaluate renewal costs based on a two-year average of payments rather than just the current year's price. In this sense, controlling for the initial joining price partially accounts for the relative price change users face at renewal, which may influence renewal decisions above and beyond the level of the renewal price itself.

In addition, user preferences regarding price and referrals may vary. For example, Apple device users may have a higher propensity to renew, possibly due to their willingness to pay premium prices for Apple products. Our data allow us to identify whether payments were made using Apple devices. We control for this variable in our model, which helps to adjust for any potential bias this factor may introduce.

Community Characteristics. Although our model includes community-level fixed effects, these are static and do not vary over time. To further account for external environmental changes that may influence users' decisions, we include several community-level time-varying characteristics in our regression model. For each user, we control for the number of articles posted and the log-transformed number of users in their community in the year prior to their renewal. We also include the total number of questions answered by community owners. These variables are proxies for the engagement of community owners and the overall vibrancy and quality of the community, which could influence users' renewal decisions. Recognizing that communities are made up of different proportions of iPhone users, we also control for the percentage of iPhone users who renew within the same time period as the user in question. A higher proportion of active iPhone users could create a more positive renewal environment, potentially influencing the renewal decisions of other users. Controlling for this variable helps us further isolate the effects of referral network and pricing on user renewal behavior.

3.2.2 Results

The results of our estimation are presented in Table 11, which outlines the relationship between various factors and users' renewal decisions. In Columns (1) to (4), we observe a significant negative correlation between the price and users' willingness to renew their subscriptions. These results indicate an inverse relationship, suggesting that higher prices have a negative impact on users' likelihood to renew in our dataset. This trend is consistent with the economical prediction: increased costs tend to reduce the propensity to repeat purchases or subscriptions.

Our analysis also reveals an unexpected trend regarding referred users. Contrary to the findings in previous literature, such as Beaman and Magruder (2012), which suggest that referred customers often exhibit higher loyalty due to a better product fit, our data indicate that referred users have a lower propensity to remain within the community. This discrepancy may be due to the specific nature of digital products and platforms. In digital contexts, where the cost of making referrals through digital channels is minimal, it is plausible that individuals are being referred who may not be an ideal fit for the community's content. As a result, these users may be less likely to renew their subscriptions.

Conversely, our results show that a user's intention to renew is positively influenced by the renewal decision of their referrer and the increase in the number of referees associated with them. This observation suggests that social interaction and communication within the community, as discussed in Van den Bulte et al. (2018), may be essential in motivating users to renew. It implies that even if the product or community is not perfectly aligned with the preferences of the referred users, the social dynamics and sense of belonging within the community may have a significant impact on their decision to renew. While these results appear to be consistent with theoretical expectations, it is important to recognize that potential endogeneity issues may still affect our findings. In the following section, we focus on these concerns and employ a series of robustness checks to mitigate their impact.

Table 11. Reduced Form Estimation Results

Dependent Variable:	*Renewal Decision* (d)				
	(1)	(2)	(3)	(4)	(5)
	OLS	OLS	OLS	OLS	IV
Price Effect					
Price	−0.098***	−0.083***	−0.068***	−0.058***	−2.134***
	(0.001)	(0.002)	(0.002)	(0.003)	(0.040)
Network Effect					
Referred	−0.245***	−0.253***	−0.180***	−0.050***	−0.044***
	(0.008)	(0.008)	(0.007)	(0.007)	(0.010)
Referrer's Decision	0.247***	0.179***	0.141***	0.030***	0.031***
	(0.009)	(0.009)	(0.008)	(0.007)	(0.012)
Number of Referees	0.254***	0.185***	0.152***	0.149***	0.123***
	(0.006)	(0.006)	(0.005)	(0.005)	(0.008)
User Characteristics					
Joining Price		0.023***	0.029***	0.008***	0.173***
		(0.002)	(0.002)	(0.002)	(0.005)

Smartphone Model		0.419***	0.319***	0.254***	0.692***
		(0.014)	(0.014)	(0.013)	(0.022)
Community Characteristics					
Number of Answers		0.005***	0.004***	0.012***	0.037***
		(0.001)	(0.001)	(0.003)	(0.004)
Number of Articles		-0.005***	-0.002	-0.014***	-0.102***
		(0.002)	(0.002)	(0.004)	(0.006)
Number of Users		0.069***	0.051***	0.006***	0.127***
		(0.001)	(0.001)	(0.002)	(0.004)
Group of iPhone Users		0.085***	-0.039***	0.100***	0.385***
		(0.015)	(0.015)	(0.014)	(0.023)
Observations	263,161	263,161	263,161	263,161	263,161
Joining Time FE	NO	NO	YES	YES	YES
Renewal Time FE	NO	NO	YES	YES	YES
Community FE	NO	NO	NO	YES	YES
R^2	0.031	0.085	0.193	0.323	-0.643

Notes: ∗p < 0.10, ∗∗p < 0.05, ∗∗∗p < 0.01. Standard errors are reported in parentheses. We use $\log(x+1)$ for the logarithmic transformation of variable x to avoid the situation where $x = 0$, and $\log(x)$ does not exist. The variable "Smartphone Model" is coded as 1 for iPhone users and 0 for non-iPhone users. It is worth noting that when using Instrumental Variable (IV) estimation, the focus is more on the validity of the instruments and the consistency of the parameter estimates, rather than the model's goodness-of-fit. In the fifth column, where IV regression is applied, we observe a negative R^2. This is because R^2 is calculated as $1 -$ (Model's Prediction Error/Prediction Error Using Mean Only). In our case, since the IV largely corrects for the price elasticity, it can lead to a prediction error that is larger than the error derived from using only the mean.

3.3 Robustness Checks

In Equation 5, we acknowledge that both network statistics and pricing may have inherent endogeneity problems. In the previous section, we addressed these issues by controlling for a wide range of variables and implementing various fixed effects. In particular, the inclusion of community-level fixed effects was intended to mitigate the impact of unobservable factors that might simultaneously influence both the referral network within a community and users' renewal decisions. However, while this set of controls and fixed effects is comprehensive, it may not fully address all endogeneity concerns.

While referral relationships between users are clearly identified in our dataset, the factors driving their decisions may correlate with aspects other than the referral relationship itself. For example, shared interests or preferences could influence both their decision to join a particular community and their subsequent renewal decision. In addition, the pricing strategy adopted by community owners typically reflects the overall renewal demand during the payment period. This "equilibrium price" could potentially bias our estimate of the true impact of price on renewal decisions. If the equilibrium price already incorporates users' response to price changes, our estimates may understate the true price elasticity. In this section, we employ various robustness checks to strengthen the reliability and credibility of our regression results.

3.3.1 Addressing Network Endogeneity

The majority of users typically refer eligible individuals within the first year, leading to an essentially predetermined network structure by the time they face renewal decisions. In addition, the referral reward system, where financial incentives are offered only during the initial referral period, allows us to isolate the effects of network structure and pricing on renewal decisions. Nevertheless, there are several alternative explanations that could potentially influence our empirical results.

Community Quality as a Driver of Renewals. First, it is possible that referrers recommend communities to referees based on the perceived quality of the community, which could influence both parties' renewal decisions. To address this, we control for community-level fixed effects. Table 11 shows that the significance and direction of our results persist even after accounting for these fixed effects. However, we observe a slight reduction in the magnitude of the influence of the referrer on renewal decisions, suggesting the importance of community quality in driving these decisions.

To further validate that the estimated effects are indeed due to relationships within the referral network and not to other unobserved factors, we employ a special strategy. This involves shuffling the referral relationships among users, thereby creating artificial network structures. If referral relationships were the sole driver of the observed network effects, the reshuffled networks would have no significant effect on the model results. Table 12 shows the results of these OLS regressions using falsified networks. Column (1) should be compared directly with Column (4) of Table 11, since it uses the same specification with the referral links replaced by falsified ones.

Unobserved Correlated Preferences. Second, unobserved correlated preferences between referrers and referees could also lead to statistical correlations. To rule out the second mechanism, our study included an extensive supplementary analysis. We hypothesized that if unobserved preferences significantly affected our results, then network effects in communities that developed primarily through private channels (e.g., direct messages on WeChat) would likely differ from those that developed primarily through public channels (e.g., web page advertisements). This assumption is based on the idea that private channel recommendations may be driven more by personal connections and shared preferences, as opposed to public channels, which may attract a more diverse user base.

Table 12. OLS Regression Results Using Falsified Networks

Dependent Variable:	*Renewal Decision* (d)		
	Original	Platform Level	Community Level
	(1)	(2)	(3)
Price Effect			
Price	−0.058***	−0.059***	−0.059***
	(0.003)	(0.003)	(0.003)
Network Effect			
Referred	−0.050***	−0.004	−0.001
	(0.007)	(0.003)	(0.005)

Referrer's Decision	0.030***	0.020	−0.003
	(0.007)	(0.014)	(0.007)
Number of Referees	0.149***	0.008	0.001
	(0.005)	(0.015)	(0.005)
User Characteristics	YES	YES	YES
Community Characteristics	YES	YES	YES
Observations	263,161	263,161	263,161
Joining Time FE	YES	YES	YES
Renewal Time FE	YES	YES	YES
Community FE	YES	YES	YES
R^2	0.323	0.320	0.320

Notes: ∗p<0.1; **p<0.05; ***p<0.01. Standard errors are reported in parentheses.

To test this hypothesis, we compared the network effects of these two different types of community growth channels. The approach involved estimating the network effects for communities with a high proportion of private channel referrals versus those with a predominance of public referral links. Our goal was to determine whether the underlying preference correlations, which may be more pronounced in private referral communities, would manifest significantly different network effects. The empirical evidence presented in Appendix P indicates that there are no significant differences in the network effects associated with these different referral modalities. This result suggests that the structure of the referral network and its effect on users' renewal decisions are not unduly influenced by unobserved correlated preferences. Such findings add robustness to our argument and reduce concerns that referral network formations may bias the results of our study.

3.3.2 Addressing Price Endogeneity

Although we find evidence that community owners rarely adjust their prices, so that the price is often unresponsive to the immediate conditions within the community, there may still be unobserved factors at play. For example, platform-level promotions or immediate incentives from the community owner could simultaneously influence both price and renewal decisions, thereby biasing our estimates. This highlights the importance of controlling for time- and community-level fixed effects and community-level time-varying characteristics. It also prompts us to introduce a set of instrumental variables (IVs) in this section to correct for potential biases in the price elasticity.

The construction of our IVs is inspired by the classic Hausman-type instruments (Hausman et al., 1997). Additionally, similar IVs have been widely used and deemed valid in marketing and information systems research (De Matos et al., 2014; Oestreicher-Singer and Sundararajan, 2012; Tucker, 2008). For each user within a community, we calculate the number of users who are up for renewal in the same period but do not have a referral relationship with her. This number is considered an exogenous shock and is used as an instrument. Our reasoning is that while users may be aware of the current follower count of the community owner (i.e., the number of users from the previous year, which has already been controlled for by community-level time-varying characteristics), the focal user is typically unaware of users with whom she does not have a referral relationship. Therefore, we believe that our instrument satisfies the exclusion restriction criterion. In addition, this subset of renewing users obviously influences the community owner's pricing strategy for the period, which also satisfies the relevance condition. Following the same logic, we use as a second instrument the average number of renewals among these users who do not have a referral relationship with them but are up for renewal in the same period. The average number of renewals among users outside a focal user's referral network is clearly information that the focal user would not have. Moreover, if the average renewal rate is high, indicating that users renewing in the same period are relatively loyal, this would significantly influence the community owner's pricing decision.

The advantage of proposing two instruments is that it allows us to statistically test whether they are weak instruments and if they can jointly pass the exclusion restriction test Goldfarb et al. (2022). The test results show that the Wu-Hausman test for weak instruments is passed at the 1% level. The first stage test for weak instruments shows significance at the 1% level. This is evidenced by a high first stage F-statistic value of 13,185.09 and a substantial partial R-squared value of 0.90, confirming the relevance of our introduced IVs. The Hansen J test for joint exclusivity of the two instruments yields a p-value of 0.718, which is not significant. It is worth noting that our sample includes almost 260,000 observations and more than 200 fixed effects; passing the Hansen J test in such circumstances suggests that the set of IVs is at least valid in our context. Column (5) of Table 11 presents the results of our IV estimation. The results indicate that the IVs significantly correct an initial underestimation of the magnitude of the price effects. With these adjustments, the data reveal an increased sensitivity of users to price changes. In particular, the coefficient of price on renewal decisions shifts significantly from -0.058 to -2.134. This adjustment leads to results that are more intuitively plausible, suggesting that our IVs effectively eliminate biases in the estimated price elasticity due to unobservable demand-side factor.

Chapter 3: Referral Networks and Renewal Rates: A Structural Model

This chapter develops and applies a structural model of renewal in referral networks. Section 1 lays out the model. Section 2 details the estimation and empirical fit. Section 3 examines uniform price changes and discount policies and Section 4 closes the chapter.

1. Structural Model Development

In this section, we propose a more sophisticated structural model to describe the renewal decisions of individual users as their subscription period approaches expiration. Unlike the Linear Probability Model (LPM) presented in Chapter 2 (Section 3), our structural model takes into account focal users' expectations about the future actions of referees within the referral network. This distinction is important because the reduced-form model is designed to document empirical correlations among prices, referral relationships, and renewal outcomes, but it does not provide a framework for modeling strategic interdependence within the network, particularly how a user's renewal decision depends on expectations about the future actions of downstream referees. By contrast, the structural model allows us to incorporate these expectations explicitly and to evaluate how changes in prices, referral policies, and network structure affect both individual renewal decisions and aggregate retention outcomes. This allows us to perform counterfactual analyses based on a nonlinear model specification to assess the effects of different referral and pricing policies.

We consider a total of M communities. For each community m, we denote $I_{m,t}$ as the set of users facing a renewal decision at time t, where $t \in \{1, \ldots, T\}$. The aggregate number of users in community m during year t is given by $n_{m,t} = \#I_{m,t}$, with $m \in \{1, \ldots, M\}$, where # signifies the cardinality of the set. A user i belonging to $I_{m,t}$ is confronted with the choice of whether to extend her membership in the community for an additional year. This decision is binary:

$d_{i,m,t} \in \{0,1\}$, where

$$d_{i,m,t} = \begin{cases} 1, & i \text{ decides to renew in year } t \\ 0, & i \text{ leaves the community in year } t \end{cases}$$

A user i will opt for $d_{i,m,t} = 1$ if the utility gained from renewing surpasses that of leaving the community.[8]

[8] We abstract away the possibility that users may be considering renewals in more than one community during the same time period. As detailed in Appendix B, only 2.21% of users hold subscriptions in more than one community. Given this small proportion, we can justifiably ignore the likelihood of users engaging in multi-homing in our further investigations.

Within a community, a user may take on the role of either a referrer or a referee. The matrix of referral relationships among users in community m at time t can be represented by a

square matrix $\mathcal{R}_{m,t} \in \mathcal{M}_{n_{m,t} \times n_{m,t}}$, where each element $r_{(i,j)} \in \{0,1\}$ is a binary indicator defined by

$$r_{i,j} = \begin{cases} 1, & \text{if } i \text{ is } j'\text{s referrer at time } t, \\ 0, & \text{otherwise.} \end{cases}$$

While each user can be referred by only a single individual, they may act as a referrer for multiple users. The cumulative number of referees for referrer i at time t is computed by $\sum_{k \in I_{m,t}} r_{i,k}$.

1.1 Random Utility Model

In our study, we treat the renewal decisions of users who renew multiple times as separate, independent events. For a user i in community m, the utility derived from opting not to renew membership at time t is denoted as $u_m\left(i, d_{i,m,t} = 0\right) = \varepsilon^0_{i,m,t}$, where $\varepsilon^0_{i,m,t}$ symbolizes the user-specific shock to the preference for non-renewal ($d_{i,m,t} = 0$). Conversely, the utility of renewing membership is given by:

$$\begin{aligned} u_m(i, d_{i,m,t} = 1) \quad = \quad & \underbrace{\beta_p \times \ln price_{i,m,t}}_{\text{Price Effect}} + \\ & \underbrace{\beta_R \times R_{i,m,t} + \beta_{RD} \times RD_{i,m,t} + \beta_{NR} \times \ln NR_{i,m,t} + \beta_r \times \mathrm{E}\left(\frac{\sum_{k \in I_{m,t}} r_{i,k} d_{k,m,t}}{\sum_{k \in I_{m,t}} r_{i,k}} \middle| d_{i,m,t} \right)}_{\text{Network Effect}} + \\ & \underbrace{\gamma_x' x_i + \gamma_m' z_{m,t}}_{\text{User and Community Characteristics}} + \underbrace{v_t + \eta_m}_{\text{Fixed Effects}} + \varepsilon^1_{i,m,t}, \end{aligned}$$

where x_t represents the vector of individual user attributes, and $z_{m,t}$ encapsulates community characteristics reflective of the community's quality at time t. The term v_t is used to denote time-level fixed effects, and η_m represents community-level fixed effects. The term $\varepsilon^1_{i,m,t}$ captures the user-specific shock to the preference for renewal ($d_{i,m,t} = 1$). These preference-specific shocks are assumed to follow an Extreme Value Type 1 distribution and are observable only to users, not econometricians. Notably, compared to the Linear Probability Model (LPM) used in Section 3.2 of Chapter 2, we incorporate a new term into the structural utility function: $\mathrm{E}\left(\sum_{k \in I_{m,t}} r_{i,k} d_{k,m,t} / \sum_{k \in I_{m,t}} r_{i,k} \,\middle|\, d_{i,m,t}\right)$, which is the expected ratio of a user's referees who renew their membership after observing the user's decision at t, within the given network structure $\mathcal{R}_{m,t}$. As the platform requires referrers to make renewal decisions prior to their referees, the coefficient β_r quantifies the impact of the referees' potential decisions on the referrer's own renewal choice.

1.2 Model Assumptions and Equilibrium

To ensure the identification of our model, we establish the following two critical assumptions:

Assumption 1. *Users are myopic, and they take the referral network $\mathcal{R}_{m,t}$ as exogenously given.*

Assumption 1 allows us to sidestep the extensive computational demands that would be involved in incorporating the anticipation of dynamic shifts in the referral network within themodel. This assumption is pragmatic given that in our dataset, over 85% of renewals occur only once and most communities have an average lifespan not exceeding three years. Hence, we infer that digital consumers typically exhibit myopic behavior, basing their renewal decisions on present conditions rather than on long-term projections. Our paper focuses on examining the interplay between renewal decisions and the existing network, rather than the formation of the network itself, which is considered an exogenous factor at the point of renewal. The decision to invite peers into the network may be influenced by the user's social ties and the platform's referral policies, but we posit that the prevailing referral policies do not directly influence renewal decisions, as the platform's incentives are triggered solely when referees join the community (as detailed in Section 2.1 of Chapter 2). The robustness checks provided in Section 3.3 of Chapter 2, support our approach of abstracting from the endogenous formation of the network in our analysis.[9]

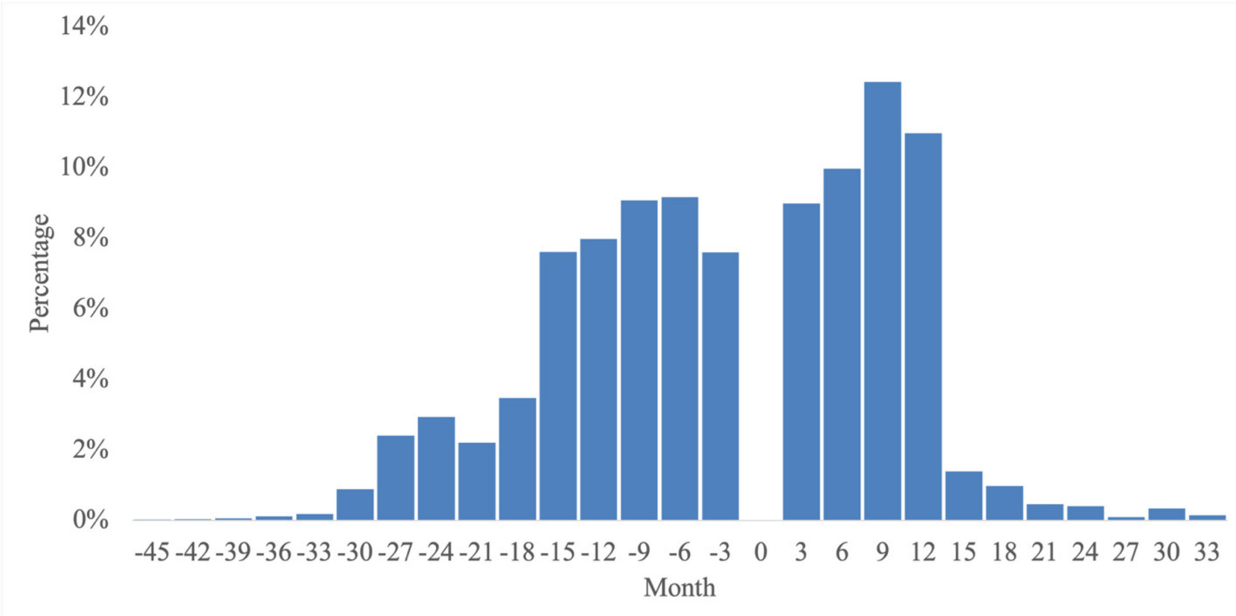
Assumption 2. *Users act independently rather than in concert, indicating the absence of collective decision-making.*

Assumption 2 assumes that users make independent decisions about leaving the community, even though their initial joining may have been influenced by trust and referral ties. This assumption is based on the understanding that there is no incentive for referrers and referees to coordinate their departure times. Such coordination would not provide any additional benefits as any bilateral agreement to coordinate departures would be unenforceable once a user has left the community. Adding to this complexity, a referrer typically recommends the community to multiple referees, making any potential coordination even more difficult and impractical.

Empirical evidence also shows that only 17% of referral-linked pairs choose to leave the community within the same month, suggesting a lack of coordinated exit strategies. Figure 5 further illustrates the different exit timelines of referrers and their respective referees. The figure shows that about half of the referees leave the community before their referrer, while the other half leave after the referrer. This distribution of exit times among referrer-associated pairs does not indicate a systematic or concerted pattern of leaving the community. The data thus reinforce

[9] In Appendix D, we present further evidence that the myopic assumption is fundamentally valid. We calculate the renewal rate for each municipality and compare it to the renewal rate for the same month in the previous year. When we regress these rates against each other and control for community-level fixed effects, we find no significant correlation between the two. This suggests that the aggregate decisions of a previous period do not influence the decisions of the current period, supporting that Assumption 1 does not affect our interpretation of the community level data.

the premise that users' decisions to leave the community are made independently, rather than as part of a collective strategy.



Notes: The figure illustrates the time gap between the departure dates of referrers and referees in a community. It uses the horizontal axis to display the time difference, measured in months, between the departure of one party (either the referrer or referee) and the other. The chart's left segment illustrates scenarios where the referrer exits the community prior to their referees (for instance, if the referee's departure was in July 2020 and the referrer's in July 2019, the time difference is recorded as -12 months). Conversely, the right segment of the graph captures instances where the referrer withdraws subsequent to their referees (for example, if the referee left in July 2019 and the referrer in July 2020, the time gap is denoted as 12 months).

Figure 5. Comparative Exit Timing of Referrers and Referees

Definition 1. For community m at time t, we define the market equilibrium as a collection of decisions $\mathcal{E}_{m,t} = \{d_{1,m,t}, \dots, d_{i,m,t}, \dots, d_{n_m,m,t}\}$ such that:

$$\forall i \in I_{m,t}, d_{i,m,t} \in \arg\max_{d\in\{0,1\}} u_m\left(i, d \middle| \mathcal{R}_{m,t}, RD_{i,m,t}, \ln NR_{i,m,t}\right).$$

Calculating the market equilibrium becomes increasingly complex as the network scales and referral connections become more complex. This complexity is exacerbated when individuals are both referrers and referees, where upstream decisions can significantly affect downstream actions. In Appendix R, we provide an approach to computing the model equilibrium.

2. Estimation and Results

We observe market equilibria $\{\mathcal{E}_{m,t}\}_{m=1,\dots,M,t=1,\dots,T}$ from the data. Each market equilibrium $\mathcal{E}_{m,t}$ is characterized by a set of decision variables $\left(d_{i,m,t}\right)_{i\in I_{m,t}}$. Within the unique network structure at time t, $\mathcal{R}_{m,t}$ a collection of individual characteristics x_i, a vector of community attributes $z_{m,t}$, and bunch of fixed effects. The structural model encapsulates a large vector of parameters

$$\theta = \left(\beta_p, \beta_R, \beta_{RD}, \beta_{NR}, \beta_r, \gamma_x', \gamma_m', v_t, \eta_m\right),$$

where γ_x and γ_m are deduced through variations in x_i and $z_{m,t}$ over time. Compared to reduced-form estimation, structural estimates based on nonlinear models are evidently more time-consuming.

We face two main challenges in estimation: First, we introduced an additional term, $\mathrm{E}\left(\sum_{k\in I_m} r_{i,k} d_k \,/\, \sum_{k\in I_m} r_{i,k} \,\middle|d_i\right)$, which cannot be directly observed and its computation is dependent on the value θ. Given that our dataset contains over 260,000 observations and the network structure within each community is incredibly complex, direct parameter estimation is almost infeasible due to the computational burden; second, the renewal price is an endogenous variable. In this section, we address each of these issues.

2.1 Estimation Procedure

The full likelihood estimation is contingent upon maximizing the function:

$$\max_{\theta} \sum_{t=1,\dots,T} \sum_{i\in I_{m,t}} \sum_{m=1,\dots,M} \ln \mathbf{P}(d_{i,m,t}|x_i, z_{m,t}, \mathcal{R}_{m,t}, \sum_{j\in I_{m,t}} r_{j,i} d_{j,t}, \mathbf{E}\left(\sum_{k\in I_{m,t}} r_{i,k} d_{k,m,t}(\theta)|d_{i,m,t} = 1\right); \theta)$$

where $\sum_{j\in I_{m,t}} r_{j,i} d_{j,m,t}$ is the referrer's decision, observable in the data.

The method for computing $\mathbf{E}\left(\sum_{k\in I_{m,t}} r_{i,k} d_{k,m,t}(\theta)|d_{i,m,t} = 1\right)$ is detailed in Appendix R, and it involves complex computations. Drawing inspiration from Su (2014), we adopt the following sequential estimation algorithm to alleviate these computational challenges.

1. We estimate an auxiliary model, which serves as our initial θ estimate. For instance, we might use:
$$\hat{\theta}_{old} \in \arg\max_{\theta} \sum_{t=1,\dots,T} \sum_{i\in I_{m,t}} \sum_{m=1,\dots,M} \ln \frac{\exp\{v_m(i, d_{i,m,t} = 1|E^d_{i,m,t} = 0)\}}{1 + \exp\{v_m\left(i, d_{i,m,t} = 1\middle|E^d_{i,m,t} = 0\right)\}}.$$
The formula approximates a classic logit model by omitting the expectation term.
2. Using the initial estimate of θ, we calculate the predicted renewal probabilities for all users $i \in I_{m,t}$ conditional on their referrer renewing (i.e., $\sum_{j\in I_{m,t}} r_{j,i} d_{j,m,t} = 1$), across all $m \in \{1, \dots, M\}$ and $t \in \{1, \dots, T\}$. We denote this probability as $p^1_{i,m,t,1}(\hat{\theta}_{old})$.
3. The predicted probability is used to compute the expected value of $\sum_{k\in I_{m,t}} r_{i,k} d_{k,m,t}(\theta)$ for $d_{i,m,t} = 1$, yielding $\hat{E}^d_{i,m,t}\left(\hat{\theta}_{old}\right) = \sum_{k\in I_{m,t}} r_{i,k} p^1_{k,m,t,1}(\hat{\theta}_{old})$.
4. With the computed expected values, $E^d_{i,m,t}s$, we update the estimate of θ by maximizing:

$$\hat{\theta}_{new} \in \arg\max_{\theta} \sum_{t=1,\ldots,T} \sum_{i \in I_{m,t}} \sum_{m=1,\ldots,M} \ln \frac{\exp\{v_m(i, d_{i,m,t} = 1 | E^d_{i,m,t} = \hat{E}^d_{i,m,t}(\hat{\theta}_{old}))\}}{1 + \exp\{v_m\left(i, d_{i,m,t} = 1 \middle| E^d_{i,m,t} = \hat{E}^d_{i,m,t}(\hat{\theta}_{old})\right)\}}.$$

5. We revise the expected probability measure using the newly estimated results, obtaining $p^1_{i,m,t,d}(\hat{\theta}_{new})$ for all i, m and t.
6. The expected value of $\sum_{k \in I_{m,t}} r_{i,k} d_{k,t}(\theta)$ given $d_{i,t} = 1$ is updated by $\hat{E}^d_{i,m,t}(\hat{\theta}_{new}) = \sum_{k \in I_{m,t}} r_{i,k} p^1_{k,m,t,1}(\hat{\theta}_{new})$.
7. We adopt the discrepancy between the previous expectations and the updated expectations as our criterion, quantified by:

$$Q(\hat{\theta}_{old}, \hat{\theta}_{new}) = \sum_{t=1,\ldots,T} \sum_{k \in I_{m,t}} \sum_{m=1,\ldots,M} \{\left(\hat{E}^d_{i,m,t}(\hat{\theta}_{old}) - \hat{E}^d_{i,m,t}(\hat{\theta}_{new})\right)^2\}.$$

We repeat Step 2 to 7 until $Q(\hat{\theta}_{old}, \hat{\theta}_{new}) < \varepsilon$, where ε is sufficiently small threshold determined by the econometrician.

Our methodology is similar to the approach described by Aguirregabiria and Mira (2007), where the convergence of the model depends on ensuring that the choice probabilities of all platform users match the expected values in the referrer's utility function. By omitting iterative processes, our approach is similar to the methods used by Dubé et al. (2010) and Ryan and Tucker (2012). These scholars estimate the best response function directly from the dataset without requiring strict consistency, a more demanding assumption than that proposed by Bajari et al. (2007).[10]

2.2 Estimation Results

In Table 13, we display the results of our structural model estimations. Columns (1) and (2) show the outcomes for logit and probit models without incorporating $\hat{E}^d_{i,m,t}$ as a control variable. Columns (3) and (4) include estimated outcomes where we consider users' expectations about the actions of referees in the subsequent period. Column (3) represents parameter estimations under incomplete information, as we do not iterate to obtain stable values of $\hat{E}^d_{i,m,t}$. Column (4), however, reflects outcomes derived from iterations, depicting the model's equilibrium under complete information. The significance of the control function term (η) across all columns, as indicated below the table, suggests that our IVs effectively corrects for price effects in both linear and non-linear specifications.

---

[10] To validate the effectiveness of our estimation method, we ran simulations for different sample sizes, as detailed in Appendix F. We estimated these samples using our approach and reported the results within 95% confidence intervals. The results consistently show that our method reliably reflects the true parameters across different sample sizes, underlining its robustness. To account for endogeneity in the price associated with renewals, we use a control function approach in estimating our structural model, following the methodology outlined by Petrin and Train (2010) and Wooldridge (2015).

Table 13. Structural Estimation Results

Dependent Variable:	*Renewal Decision* (d)			
	Logit	Probit	Incomplete Information	Complete Information
	(1)	(2)	(3)	(4)
Price Effect				
Price	−1.435***	−0.814***	−1.435***	−1.435***
	(0.038)	(0.022)	(0.038)	(0.038)
Network Effect				
Referred	−0.250***	−0.141***	−0.249***	−0.249***
	(0.041)	(0.024)	(0.041)	(0.041)
Referrer's Decision	0.137***	0.073***	0.132***	0.133***
	(0.045)	(0.026)	(0.045)	(0.045)
Number of Referees	1.486***	0.801***	0.636***	0.721***
	(0.047)	(0.026)	(0.072)	(0.073)
Expected of Referees' Decisions			1.575***	1.384***
			(0.117)	(0.115)
User Characteristics	YES	YES	YES	YES
Community Characteristics	YES	YES	YES	YES
Control Function				
$\hat{\eta}$	0.004***	0.004***	0.004***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	263,161	263,161	263,161	263,161
Joining Time FE	YES	YES	YES	YES
Renewal Time FE	YES	YES	YES	YES
Community FE	YES	YES	YES	YES
Log Likelihood	-129,472.900	-129,491.100	-129,372.600	-129,393.900

Notes: ∗p<0.1; **p<0.05; ***p<0.01. Standard errors are reported in parentheses.

The estimation results in all columns are consistent with those in Table 11, confirming the stability of the price and network effects estimates when controlling for a sufficient number of variables. Comparing Columns (1) and (2) with the others, we observe that the inclusion of expectations improves the fit of the model, as indicated by an increased likelihood. Additional tests confirm that the model with prediction elements ($\hat{E}^{d}_{i,m,t}$) statistically outperforms the traditional logit model at the 1% significance level. In particular, Columns (3) and (4) show no significant differences between the estimates under complete and incomplete information scenarios. The results suggest that a 1% price reduction is correlated with a 1.435% increase in the marginal propensity to renew. The consistency of network effects across different model specifications can be attributed to the demonstrated exogeneity of network-related variables (as discussed in Section 3.3 of Chapter 2) and the fact that the IVs are constructed based on users outside the focal user's referral network, thus not affecting the estimation of network effects.[11]

3. Policy Simulations

The results of our structural model suggest that referral relationships directly influence users' renewal decisions, and that the effect of price changes on renewal rates varies according to different network structures within each community. One of the key questions for both academic and business practitioners is how the effects of discounting propagate through these referral networks, and which specific individuals should be targeted with discounts to encourage renewals. We conduct counterfactual analyses to explore these issues.

3.1 Uniform Price Changes

We begin by examining the effects of pricing changes, focusing on the effect of a discounting strategy on renewal prices and average renewal rates. In practice, the platform allows community owners to set renewal prices independently. For the purposes of our analysis, we now allow the platform to implement a platform-wide promotional campaign. This campaign would enforce a uniform renewal discount across all communities, offering a percentage discount to all renewing users on the platform.We define this discount as "disc" in our counterfactual study.

First, we need to develop an algorithm to calculate the equilibrium response to discounts, since price changes affect renewal probabilities throughout the network. Our process for calculating the counterfactual equilibrium under a price discount policy involves several steps:

1. Modify the model outcomes by setting $price_{i,m,t}^{CF} = price_{i,m,t} \times (1 - disc)$, where $disc$ represents the proportional price discount enacted by the platform. We then calculate a set of model-based renewal probabilities.
2. With an array of preference shocks, recalibrate the market equilibrium to

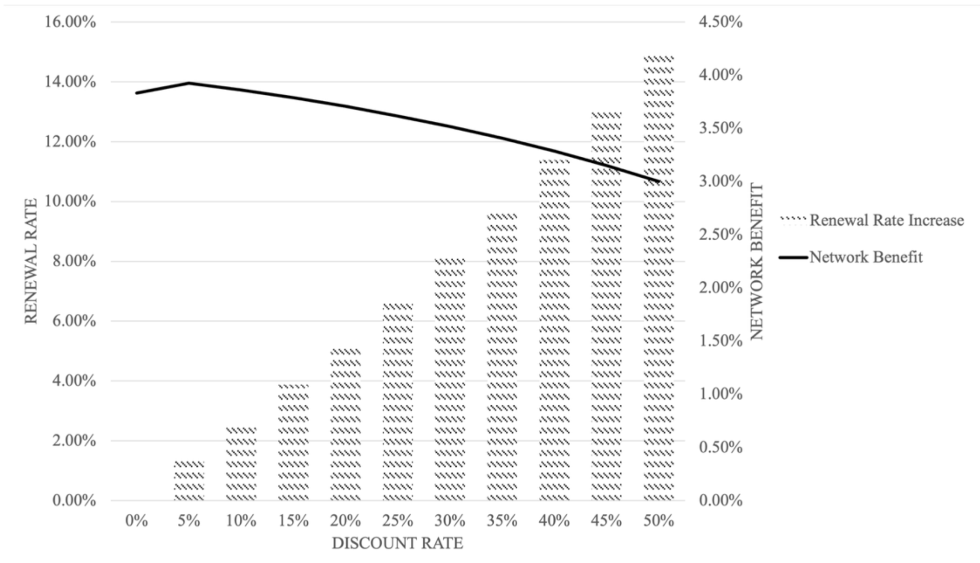
$$\mathcal{E}_{m,t}^{CF} = \{d_{1,m,t}, \dots, d_{i,m,t}, \dots, d_{n_m,m,t}\};$$

3. Update the referrer's status according to $\mathcal{E}_{m,t}^{CF}$, and refresh the value of $\hat{E}_{i,m,t}^{d}$s based on the model-fitted probabilities.
4. Reassess the equilibria and repeatedly apply Steps 2 and 3 until both $\mathcal{E}_{m,t}^{CF}$ and $\hat{E}_{i,m,t}^{d}$s stabilize.

In this study, we assume that users make renewal decisions myopically, treating each renewal as an isolated event. Such an assumption suggests that our counterfactual analysis may slightly

[11] In Appendix G, we report the average marginal effects corresponding to the coefficients estimated via the logit and probit models. Despite the differences in the direct coefficient estimates, the associated average marginal effects are highly consistent, reinforcing the reliability of our findings.

underestimate the aggregate effect of user departures, as leaving the community is not considered to have a lasting effect in any subsequent period.



Notes: Our counterfactual analysis is based on a group of users with network relationships (i.e. each user has at least one referrer or referee), which represents about 10% of the total number of users. The horizontal axis represents the change in total discount. The dashed bars represent the increase in the renewal rate as the price falls, and the dark line measures the effect of the network by comparing the difference in renewal rates with and without the referral network.

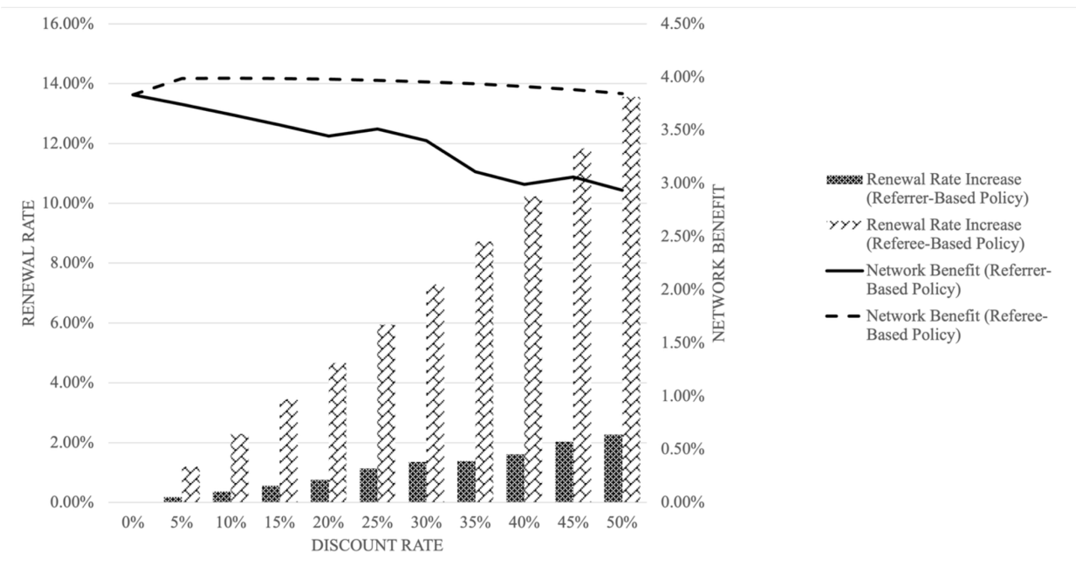
Figure 6. Price Discount, Renewal Rate and Network Benefit

Figure 6 illustrates the effect of price changes. Reducing prices increases the likelihood that users will renew their subscriptions and changes their renewal expectations. A substantial discount, such as 50%, leads to a significant increase in total renewal rates of 15%, which is in line with the average marginal effects discussed in Appendix T. The figure provides an empirical basis for developing discounting strategies, as platforms often balance direct user revenue against advertising revenue. Due to the relatively low price elasticity of digital content subscriptions at the renewal stage, universal discounts could paradoxically reduce immediate total revenue. Our results suggest that a 1% increase in renewal rates could potentially lead to a 0.64% decrease in total revenue.

Our analysis also explores the effect of network spillover on renewal rates. We examine the consequences of network disconnection, where users disregard the renewal decisions of their referrers and referees. The network spillover benefit has a concave shape, peaking at price reductions between 5% and 10%, and increasing the renewal rate by an additional 4%. This result reflects a nuanced interplay between pricing and network effects. At lower levels of discount, the decision to renew is primarily influenced by the user's direct network contacts. As the discount increases, more users are affected and the network effect intensifies. However, for larger discounts, the decision to renew is primarily driven by the discounted price rather than by changes in network peers.

3.2 Price Discount Policies

Implementing a uniform price discount policy, while beneficial in increasing renewal rates, can adversely impact revenue. In many business models, referrals are often enticed with a one-time discount on their initial purchase. For example, it is common in the banking sector to waive the first year's annual fee for referred credit card holders. Drawing from the insights of Hinz et al. (2011), we explore the effects of price discount policies that target either referees or referrers exclusively in a distinct counterfactual scenario. Our estimations suggest that referrers may not always be well-suited to the community they join, thus highlighting them as potential targets for price-centric strategies. Offering discounts specifically to either referrers or referees could alleviate the revenue losses associated with widespread discounting and might also amplify the effectiveness of network effects. With this theory in mind, we explore a strategic pricing approach, applying discounts to one specific group within the referral network.



Notes: Our counterfactual analysis is based on a group of users with network relationships (i.e. each user has at least one referrer or referee), which represents about 10% of the total number of users. The horizontal axis represents the change in price discount.

Figure 7. Renewal Rate and Network Benefit Changes

Our counterfactual analysis examines scenarios where discounts are offered exclusively to either referees or referrers. The results, detailed in Figure 7, indicate that providing discounts to referees yields greater benefits compared to targeting referrers. This finding is consistent with the outcomes of prior field experiments (e.g., Jung et al., 2021). The observed increase in renewal rates is primarily due to the discounts given to referees. Typically, the number of referees exceeds the number of referrers, and many referees also act as referrers. As a result, discounting for referees is a more effective approach, potentially maintaining high levels of revenue due to network effects, estimated to be around 4%. Discounting for referees is also conceptually more effective. While discounting for one referrer only affects his decision, a change in the decision of one referrer can affect the renewal decisions of all his referees. On the other hand, offering discounts to referees not only changes their own renewal decisions, which affects the expectations of the referrer, but also increases the number of referees who stay in the community, which has a double effect on the network. This double margin is evident in the changes in renewal rates and network benefits shown in Figure 7.

4. Conclusions

While the referral model has traditionally been viewed as profitable but costly, the advent of digital platforms has increased its profitability and reduced its associated costs. The network effect fueled by frequent referrals not only increases the platform's revenue, but also allows for the mapping of user relationships through referral networks. In our study, we leveraged over 260,000 data points from a creator platform to describe user renewal behavior within a structural model. By examining the extensive and diverse referral networks in which users make decisions, we can identify the influence of network structure on users' strategic behavior.

The evidence from both reduced-form and structural estimations shows that referral relationships significantly affect renewal rates for digital products. Price changes ripple through the network, influencing users' expectations about their network peers' renewal decisions. These findings imply that companies and platforms need to consider the network effect of users when designing pricing or rebate strategies due to the significant differential effects across the network. We find that implementing peer-targeted discounting policies is a cost-effective way to improve renewal rates, a tactic often overlooked by practitioners in their business models.

Our study provides insightful managerial applications for platforms with different referral policies:

Platforms often design their referral networks to balance centrality and connectivity, which is influenced by the nature of their referral policies. For example, Lingoda, a prominent online language learning platform, encourages personal referrals among friends and family. They offer a €50 discount for the first month of a friend's referral and provide referrals with free credits for group classes. Importantly, Lingoda has a policy against publicly sharing referral links on coupon or review sites, a strategy that likely limits network centralization and encourages more personal, one-to-one referrals.[12] Similarly, Duolingo, another language learning platform, allows sharing of referral links on social media, but limits the number of referees to 24 per referrer.[13] Beyond language platforms, OpenAI's artificial intelligence chatbot program, ChatGPT, also shifted its focus to a low-centrality referral policy after rapidly accumulating a user base in its early stages. They limit each user to no more than three referral links to recommend others to use the paid version of the service. In addition, each referral link can only be used by one other user.[14] Limiting the number of referees effectively reduces the degree of centralization of the

[12] Source: https://lingoda-students.elevio.help/en/articles/250-can-i-recommend-lingoda-to-my-friends, accessed Nov 2022.
[13] Source: https://support.duolingo.com/hc/en-us/articles/4404225309581-How-does-the-referral-program-work-, accessed Nov 2022.
[14] Source: https://help.openai.com/en/articles/8381046-free-trial-invites-faq, accessed Nov 2023.

network by limiting the extent to which a single user can influence the structure of the network. Policies like these play a critical role in limiting the influence of super-influencers, thereby encouraging the development of less centralized networks.

Our study suggests that platforms focused on long-term growth, such as Lingoda and Duolingo, tend to favor less centralized and more loosely connected referral networks. For these platforms, increasing user retention is critical to sustained profitability. Accordingly, they design referral programs that encourage personal, one-to-one referrals and foster less centralized networks. Conversely, platforms looking for quick wins or to bolster their funding rounds often rely heavily on a few super-influencers. This strategy results in the creation of more centralized and highly connected networks. These platforms tend to have shorter operational lifespans and may not prioritize user retention, as their primary goal is to quickly grow their user base and secure initial financial commitments. As a result, their referral programs are designed to maximize connectivity, often at the expense of creating a highly centralized network structure, as they prioritize immediate user acquisition over the long-term stability provided by a more decentralized referral network. Additional analysis and anecdotal evidence supporting these implications is provided in Appendix U.

Our study also offers new management insights for governments and platforms that primarily use promotional pricing strategies as a key channel for user retention. In traditional platforms and organizations, promotional pricing has been a critical strategy for customer retention (Lewis, 2004). Our results show for the first time that pricing policies can interact with the effects of referral networks. Not only did we find that price-based loyalty can be influenced by the referral network, affecting the loyalty of both the referrer and the referee, but also that in certain scenarios the effects of referrer and referee loyalty can outweigh the effects of price reductions. Such insights are critical for platform management in the digital age, as the cost of user retention has become more expensive compared to traditional businesses.

Our study suggests that ignoring the effect induced by the referral network may overestimate the true effectiveness of promotional pricing in increasing user loyalty on digital platforms. Moreover, the results encourage platforms to more carefully consider the application of promotional pricing strategies to increase loyalty in different contexts. For example, in a highly centralized referral network, it may be more cost-effective to retain core users through discriminatory pricing than to offer renewal price incentives to all users. Conversely, in a network with low centrality but strong user ties, applying uniform price reductions to all users may maximize the effect of the network. Given the network structure, the development of differentiated, discriminatory promotion strategies should be an important direction for companies and platforms in their future user retention efforts. Properly formulated strategies can not only save costs and increase profits, but also be more conducive to long-term business development.

References

Ackerberg, D. A. and Gowrisankaran, G. (2006). Quantifying equilibrium network externalities in the ACH banking industry. *The RAND Journal of Economics*, 37(3):738-761.

Aggarwal, S. (2018). Do rural roads create pathways out of poverty? evidence from india. *Journal of Development Economics*, 133:375–395.

Aguirregabiria, V. and Mira, P. (2007). Sequential estimation of dynamic discrete games. *Econometrica*, 75(1):1-53.

Aker, J. C. (2010). Information from markets near and far: Mobile phones and agricultural markets in niger. *American Economic Journal: Applied Economics*, 2(3):46–59.

Alchemer (2021). 2021 Mobile App engagement benchmark report. Research Report. Accessed: Feburary 2024.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Ansari, A., Stahl, F., Heitmann, M., and Bremer, L. (2018). Building a social network for success. *Journal of Marketing Research*, 55(3):321-338.

Aral, S. and Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623-1639.

Armelini, G., Barrot, C., and Becker, J. U. (2015). Referral programs, customer value, and the relevance of dyadic characteristics. *International Journal of Research in Marketing*, 32(4):449-452.

Atkin, D., Khandelwal, A. K., and Osman, A. (2017). Exporting and firm performance: Evidence from a randomized experiment. *The Quarterly Journal of Economics*, 132(2):551–615.

Bahia, K., Castells, P., Cruz, G., Masaki, T., Pedr´os, X., Pfutze, T., Rodr´ıguez-Castel´an, C., and Winkler, H. (2024). The welfare effects of mobile broadband internet: Evidence from nigeria. *Journal of Development Economics*, 170:103314.

Bailey, M., Johnston, D., Kuchler, T., Stroebel, J., and Wong, A. (2022). Peer effects in product adoption. *American Economic Journal: Applied Economics*, 14(3):488-526.

Bajari, P., Benkard, C. L., and Levin, J. (2007). Estimating dynamic models of imperfect competition. *Econometrica*, 75(5):1331-1370.

Beaman, L. and Magruder, J. (2012). Who gets the job referral? Evidence from a social networks experiment. *American Economic Review*, 102(7):3574-3593.

Ben Rhouma, T. and Zaccour, G. (2018). Optimal marketing strategies for the acquisition and retention of service subscribers. *Management Science*, 64(6):2609-2627.

Belo, R. and Li, T. (2022). Social referral programs for freemium platforms. *Management Science*, 68(12):8933-8962.

Berger, T., Chen, C., and Frey, C. B. (2018). Drivers of disruption? estimating the uber effect. *European Economic Review*, 110:197–210.

Bergquist, L. F. and Dinerstein, M. (2020). Competition and entry in agricultural markets: Experimental evidence from kenya. *American Economic Review*, 110(12):3705–3747.

Bhattacharya, P., Phan, T. Q., Bai, X., and Airoldi, E. M. (2019). A coevolution model of network structure and user behavior: The case of content generation in online social networks. *Information Systems Research*, 30(1):117-132.

Biyalogorsky, E., Gerstner, E., and Libai, B. (2001). Customer referral management: Optimal reward programs. *Marketing Science*, 20(1):82-95.

Bolton, R. N., Kannan, P. K., and Bramlett, M. D. (2000). Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the Academy of Marketing Science*, 28(1):95-108.
Boudreau, K. J., Jeppesen, L. B., and Miric, M. (2022). Competing on freemium: Digital competition with network effects. *Strategic Management Journal*, 43(7):1374-1401.
Brown, M., Setren, E., and Topa, G. (2016). Do informal referrals lead to better matches? Evidence from a firm's employee referral system. *Journal of Labor Economics*, 34(1):161-209.
Brynjolfsson, E., Hu, Y., and Smith, M. D. (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, 49(11):1580–1596.
Bursztyn, L., Ederer, F., Ferman, B., and Yuchtman, N. (2014). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica*, 82(4):1273-1301.
Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.
Campbell, C., Ferraro, C., and Sands, S. (2014). Segmenting consumer reactions to social network marketing. *European Journal of Marketing*, 48(3/4):432-452.
Cao, G., Jin, G. Z., Weng, X., and Zhou, L.-A. (2021). Market-expanding or market-stealing? competition with network effects in bike-sharing. *The RAND Journal of Economics*, 52(4):778–814.
Cennamo, C. and Santalo, J. (2013). Platform competition: Strategic trade-offs in platform markets. *Strategic Management Journal*, 34(11):1331–1350.
Chu, J. and Manchanda, P. (2016). Quantifying cross and direct network effects in online consumer-to-consumer platforms. *Marketing Science*, 35(6):870-893.
Couture, V., Faber, B., Gu, Y., and Liu, L. (2021). Connecting the countryside via e-commerce: Evidence from china. *American Economic Review: Insights*, 3(1):35–50.
De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996.
De Matos, M. G., Ferreira, P., and Krackhardt, D. (2014). Peer influence in the diffusion of iPhone 3G over a large social network. *Mis Quarterly*, 38(4):1103-1134.
Demir, B., Javorcik, B., Michalski, T. K., and Ors, E. (2024). Financial constraints and propagation of shocks in production networks. *Review of Economics and Statistics*, 106(2):437-454.
Donaldson, D. (2018). Railroads of the raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5):899–934.
Dubé, J. P. H., Hitsch, G. J., and Chintagunta, P. K. (2010). Tipping and concentration in markets with indirect network effects. *Marketing Science*, 29(2):216-249.
Fernandes, A. M., Mattoo, A., Nguyen, H., and Schiffbauer, M. (2019). The internet and chinese exports in the pre-ali baba era. *Journal of Development Economics*, 138:57–76.
Fernández-Loría, C., Cohen, M. C., and Ghose, A. (2023). Evolution of referrals over customers' life cycle: Evidence from a ride-sharing platform. *Information Systems Research*, 34(2):698-720.
Fornell, C. and Wernerfelt, B. (1987). Defensive marketing strategy by customer complaint management: a theoretical analysis. *Journal of Marketing Research*, 24(4), 337-346.

Foster, A. D. and Rosenzweig, M. R. (1995). Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy*, 103(6):1176–1209.

Foster, A. D. and Rosenzweig, M. R. (2004). Agricultural productivity growth, rural economic diversity, and economic reforms: India, 1970–2000. *Economic Development and Cultural Change*, 52(3):509–542.

Galbreth, M. R., Ghosh, B., and Shor, M. (2012). Social sharing of information goods: Implications for pricing and profits. *Marketing Science*, 31(4):603-620.

Garnefeld, I., Eggert, A., Helm, S. V., and Tax, S. S. (2013). Growing existing customers' revenue streams through customer referral programs. *Journal of Marketing*, 77(4):17-32.

Goldfarb, A. and Tucker, C. (2019). Digital economics. *Journal of Economic Literature*, 57(1):3-43.

Goldfarb, A., Tucker, C., and Wang, Y. (2022). Conducting research in marketing with quasi-experiments. *Journal of Marketing*, 86(3):1-20.

Goyal, A. (2010). Information, direct access to farmers, and rural market performance in central india. *American Economic Journal: Applied Economics*, 2(3):22–45.

Hada, M., Grewal, R., and Lilien, G. L. (2014). Supplier-selected referrals. *Journal of Marketing*, 78(2):34-51.

Hartmann, W. R., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D., ... and Tucker, C. (2008). Modeling social interactions: Identification, empirical methods and policy implications. *Marketing Letters*, 19(3):287-304.

Hausman, J. A., Pakes, A., and Rosston, G. L. (1997). Valuing the effect of regulation on new services in telecommunications. Brookings papers on economic activity. *Microeconomics*, 1997:1-54.

Hildebrand, C., Häubl, G., Herrmann, A., and Landwehr, J. R. (2013). When social media can be bad for you: Community feedback stifles consumer creativity and reduces satisfaction with self-designed products. *Information Systems Research*, 24(1):14-29.

Hinz, O., Schulze, C., and Takac, C. (2014). New product adoption in social networks: Why direction matters. *Journal of Business Research*, 67(1):2836-2844.

Hinz, O., Skiera, B., Barrot, C., and Becker, J. U. (2011). Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, 75(6):55-71.

Hong, Y., Pavlou, P. A., Shi, N., and Wang, K. (2017). On the role of fairness and social distance in designing effective social referral systems. *Mis Quarterly*, 41(3):787-A13.

Hristakeva, S. and Mortimer, J. H. (2023). Price dispersion and legacy discounts in the national television advertising market. *Marketing Science*, 42(6):1162-1183.

Hu, M. M., Yang, S., and Xu, D. Y. (2019). Understanding the social learning effect in contagious switching behavior. *Management Science*, 65(10):4771-4794.

Huang, H., Huang, Y., Yan, Z., and Zhang, H. (2022). Social influence, competition, and free riding: examining seller interactions within an online social network. *MIS Quarterly*, 46(3):1817-1832.

Inkpen, A. C. and Tsang, E. W. (2005). Social capital, networks, and knowledge transfer. *Academy of Management Review*, 30(1):146-165.

Intelligence, I. (2024). Worldwide retail e-commerce forecast 2024. Accessed: July 2024.

Jakiela, P. (2021). Simple diagnostics for two-way fixed effects.

Jensen, R. (2007). The digital provide: Information (technology), market performance, and welfare in the south indian fisheries sector. *The Quarterly Journal of Economics*, 122(3):879–924.

Jin, G. Z. and Kato, A. (2006). Price, quality, and reputation: Evidence from an online field experiment. *The RAND Journal of Economics*, 37(4):983–1005.

Johnson, G. A., Lewis, R. A., and Reiley, D. H. (2017). When less is more: Data and power in advertising experiments. *Marketing Science*, 36(1):43–53.

Jung, J., Bapna, R., Gupta, A., and Sen, S. (2021). Impact of incentive mechanism in online referral programs: evidence from randomized field experiments. *Journal of Management Information Systems*, 38(1):59-81.

Jung, J., Bapna, R., Golden, J. M., and Sun, T. (2020). Words matter! Toward a prosocial call-to-action for online referral: Evidence from two field experiments. *Information Systems Research*, 31(1):16-36.

Kamada, Y. and Öry, A. (2020). Contracting with word-of-mouth management. *Management Science*, 66(11):5094-5107.

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., and Nerur, S. (2018). Advances in social media research: Past, present and future. *Information Systems Frontiers*, 20(3):531-558.

Katz, M. L. and Shapiro, C. (1994). Systems competition and network effects. *Journal of Economic Perspectives*, 8(2):93–115.

Kornish, L. J. and Li, Q. (2010). Optimal referral bonuses with asymmetric information: Firm-offered and interpersonal incentives. *Marketing Science*, 29(1):108-121.

Leenheer, J., Van Heerde, H. J., Bijmolt, T. H., and Smidts, A. (2007). Do loyalty programs really enhance behavioral loyalty? An empirical analysis accounting for self-selecting members. *International Journal of Research in Marketing*, 24(1):31-47.

Lewis, M. (2004). The influence of loyalty programs and short-term promotions on customer retention. *Journal of Marketing Research*, 41(3):281-292.

Li, C., Zheng, Y., and Lv, X. (2025). Can e-commerce alleviate household financial vulnerability? *Economic Analysis and Policy*, 87:2043–2058.

Li, P., Lu, Y., and Wang, J. (2016). Does flattening government improve economic performance? evidence from china. *Journal of Development Economics*, 123:18–37.

Lobel, I., Sadler, E., and Varshney, L. R. (2017). Customer referral incentives and social media. *Management Science*, 63(10):3514-3529.

Ma, B., Zhang, C., Guo, J., and Zhang, C. (2023). Does e-commerce increase the consumption of rural households?—a quasi-experiment of "national rural e-commerce comprehensive demonstration policy". *China Economic Quarterly*, 23(5):1846–1864.

Mandavia, M. (2022). How india plans to reinvent e-commerce. *The Wall Street Journal*.

Max Wei, Y. (2020). The similarity network of motion pictures. *Management Science*, 66(4):1647-1671.

Mele, A. (2022). A structural model of homophily and clustering in social networks. *Journal of Business & Economic Statistics*, 40(3):1377-1389.

Muller, E. and Peres, R. (2019). The effect of social networks structure on innovation performance: A review and directions for research. *International Journal of Research in Marketing*, 36(1):3-19.

Newman, D. (2015). *Love it or hate it: Influencer marketing works*. Forbes.

Oestreicher-Singer, G. and Sundararajan, A. (2012). The visible hand? Demand effects of recommendation networks in electronic markets. *Management Science*, 58(11):1963-1981.
Pallais, A. and Sands, E. G. (2016). Why the referential treatment? Evidence from field experiments on referrals. *Journal of Political Economy*, 124(6):1793-1828.
Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, 47(1):3-13.
Reingen, P. H. and Kernan, J. B. (1986). Analysis of referral networks in marketing: Methods and illustration. *Journal of Marketing Research*, 23(4):370-378.
Reshef, O. (2023). Smaller slices of a growing pie: The effects of entry in platform markets. *American Economic Journal: Microeconomics*, 15(4):183–207.
Roelens, I., Baecke, P., and Benoit, D. F. (2016). Identifying influencers in a social network: The value of real referral data. *Decision Support Systems*, 91:25-36.
Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*.
Ryan, S. P. and Tucker, C. (2012). Heterogeneity and the dynamics of technology adoption. *Quantitative Marketing and Economics*, 10(1):63-109.
Ryu, G. and Feick, L. (2007). A penny for your thoughts: Referral reward programs and referral likelihood. *Journal of Marketing*, 71(1):84-94.
Saeedi, M. (2019). Reputation and adverse selection: Theory and evidence from ebay. *The RAND Journal of Economics*, 50(4):822–853.
Schmitt, P., Skiera, B., and Van den Bulte, C. (2011). Referral programs and customer value. *Journal of Marketing*, 75(1):46-59.
Shapiro, C. and Varian, H. R. (1999). I*nformation Rules: A Strategic Guide to the Network Economy*. Harvard Business Press.
Shriver, S. K., Nair, H. S., and Hofstetter, R. (2013). Social ties and user-generated content: Evidence from an online social network. *Management Science*, 59(6):1425-1443.
Sorensen, E. (2022). *OTT: Perception, use, and business models*. Research Report, Parks Associates. Accessed: Feburary 2025.
Su, C. L. (2014). Estimating discrete-choice games of incomplete information: Simple static examples. *Quantitative Marketing and Economics*, 12(2):167-207.
Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.
Sun, T., Viswanathan, S., and Zheleva, E. (2021). Creating social contagion through firm-mediated message design: Evidence from a randomized field experiment. *Management Science*, 67(2):808-827.
Tucker, C. (2008). Identifying formal and informal influence in technology adoption with network externalities. *Management Science*, 54(12):2024-2038.
Van den Bulte, C., Bayer, E., Skiera, B., and Schmitt, P. (2018). How customer referral programs turn social capital into economic capital. *Journal of Marketing Research*, 55(1):132-146.
Vidal, M. and Faz, X. (2020). E-commerce is taking off in rural china: 3 lessons for other countries. *CGAP Blog*. Accessed: July 2025.
Wang, J. (2013). The economic impact of special economic zones: Evidence from Chinese municipalities. *Journal of Development Economics*, 101:133–147.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2):420-445.

World Bank (2019). E-commerce development: Experience from china. Technical Report Report No. 144689-CN, World Bank Group, Washington, DC. Accessed: July 2025.

Yunnan Provincial Department of Natural Resources (2024). Regional public brands: “jingmai mountain” and “pu’er jingmai mountain tea”. Official website. Chinese official website; accessed: August 2025.

Zervas, G., Proserpio, D., and Byers, J. W. (2017). The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Journal of Marketing Research*, 54(5):687–705.

Appendix A: List of GEPs in China from 2017 to 2023

Table A.1 offers a current and detailed list of China's regions where local governments have established e-commerce public service centers between 2017 and 2023. The table shows the variety of local efforts and illustrates the widespread government-backed rural digitization initiatives across China. The listed platforms mainly serve as public service centers to help local producers access and use existing commercial e-commerce marketplaces, rather than functioning as independent, profit-driven platforms. The table highlights the government's focus on improving rural digital infrastructure, branding support, and training services, reflecting the broader goals of the National Rural E-commerce Comprehensive Demonstration Program launched in 2014 and significantly expanded in 2017 (Ma et al., 2023; Li et al., 2025).

Table A.1. Government E-commerce Public Service Centers in China since 2017

Province	Location	Setup Year	Name of Public E-commerce Service Center	Link
Tianjin	Tianjin	2018	Tianjin Yinonghe	http://tj.365960.cn/catalog/11000000.html
Gansu	Minle	2019	Min Le You Pin	https://mldszx.com.cn/productlist.php
Henan	Xixian	2019	Xi County E-commerce Public Service Center	hhttp://www.xixiandianshang.cn/index.html
Gansu	Tanchang	2018	E-commerce of Tanchang	http://www.tanchangds.com/index.jsp
Shanxi	Xixian	2017	Xixian County E-commerce Public Service Center	http://xxdsfwzx.com/techan.aspx?ClassID=78
Chongqing	Qijiang District	2018	Chongqing Qijiang Caiba Trade Co., Ltd.	http://www.cb023.com/
Liaoning	Tieling	2021	Tieling E-commerce Public Service Center	http://data.ehoneycomb.net/data/index/index/city_id/13.html
Jiangsu	Jiangyin	2019	Jiangyin E-commerce Public Service Platform	http://jiangyinds.com/product
Anhui	Laian	2017	Laian E-commerce Public Service Center	http://www.laecps.com/list-teseguan-2.html
Fujian	Quanzhou	2018	Anxi E-commerce Public Service Center	http://www.axswfj.com/
Shandong	Linyi	2022	Lanling Electronic Commerce Public Service Center	http://www.lanlingds.com/
Henan	Zhoukou	2019	Shangshui E-commerce Public Service Center	http://shop.shangshui.agdata.cn/wssc.html
Guangdong	Chaozhou	2018	Chaozhou E-commerce Public Service Center	https://www.czecc.com/index.php/commerce/shop.html
Hainan	Hainan	2018	Hainan Rural Revitalization Network	https://shop.hainanfp.com/index
Heilongjiang	Jiamusi	2017	Jiamusi Specialty Website	http://jiamusi.kuaimicheng.com/techan.html
Hunan	Xiangyin	2016	Xiangyin County E-commerce Public Service Center	https://xiangyin.hnbotong.net/goods/all?page=2
Guizhou	Guiyang	2022	Kaiyang County E-commerce Public Service Center	http://www.seonky.cn/?product/
Shannxi	Yulin	2018	Qingjian County E-commerce Public Service Center	https://www.91jindi.com/index.php?homepage=15667062990&file=sell
Qinghai	Maduo	2021	Maduoxian Commerce Public Service Center	https://www.maduodianshang.com/
Inner Mongolia	Ordos	2020	Wushen Banner E-commerce Public Service Center	https://wsq.we1010.cn/specialty.html
Guangdong	Longmen	2018	Longmen County E-commerce Public Service Center	http://longmen.hunge.vip/goods
Qinghai	Huzhu	2021	Huzhu E-commerce Public Service Center	http://www.huzhuds.com/specialty?tabIndex=1
Guangxi	Shangsi	2017	Shangsi E-commerce Public Service Center	http://www.ssdszx.com/pr.jsp?_pp=0_318_0_-1&pcp=2
Xinjiang	Hetian	2021	Hetianyuese	http://hetian.pandahigo.com/

Notes: The above table lists government-initiated e-commerce public service centers identified through our research, reflecting the broad implementation of China's National Rural E-commerce Comprehensive Demonstration Program. These centers primarily assist small-scale producers in rural areas in accessing and effectively utilizing existing commercial e-commerce platforms. All listed website links were accessible as of June 2025. Additional centers may exist beyond those listed here, considering the ongoing and expanding nature of government initiatives.

Appendix B: Comparative Institutional Context: Entry Costs and Operational Complexity

The Lancang GEP differs structurally from major commercial e-commerce platforms and social media commerce channels. Two factors are important for smallholder adoption and policy design: (i) financial entry costs and (ii) operational complexity and required skills, comparing each with official rulebooks and program manuals. Throughout, we rely on platform rulebooks and large-agency reports rather than trade blogs to benchmark costs and frictions.[15]

B.1 Financial Entry Barriers

Large platforms usually require refundable deposits, platform fees, and per-transaction service fees. In Tmall Global, the official rulebook specifies a refundable security deposit and an annual technical service fee with two tiers (30,000 or 60,000 RMB). In addition, there are technical service fees per transaction that generally range from 2% to 5%, depending on the category (Tmall Global, 2024). JD Worldwide applies a flat transaction service fee of 0.9% to POP merchants and utilizes a tiered deposit scheme that increases with sales and category (JD Worldwide, 2025a,b). In contrast, Taobao (C2C) requires a refundable consumer protection deposit, the amount of which depends on the category under the Consumer Protection Service Agreement (Taobao, 2024). Beyond formal fees, participating in these platforms typically involves ongoing expenses for paid traffic and promotions. Evidence from the Taobao Village study by the World Bank highlights high advertising and promotion costs, intense competition, and lack of skills as the main challenges faced by E-shop owners (World Bank, 2019).

Short videos and social media platforms like Douyin have low formal access fees but depend on creator interaction. The official Douyin rules show (i) technical service fees for the platform by category, usually ranging from 1 to 5%, and (ii) affiliate commissions from merchants within its Jingxuan Alliance: 1 to 50% for general plans, with higher caps (up to 80%) under targeted plans (Douyin E-commerce, 2025a,b). In practice, gaining significant visibility often requires paid advertising and creator commissions, making indirect costs substantial even when headline fees are low.

In our setting, by design, the Lancang GEP does not impose deposits, listing fees, or commissions on local farmers (according to government policy and our fieldwork protocols). Public finance and screening replace monetary entry screens, reducing barriers for small-holders in low and middle-income countries (LMICs) and shaping the empirical patterns we study.

B.2 Operational Complexity and Required E-commerce Skills

Operating stores on large commercial platforms such as Tmall, JD, and Taobao requires comprehensive skills in merchandising, customer service, fulfillment, promotion, and data-driven operations. On social media platforms such as Douyin, content creation, live stream

[15] Fee schedules vary by category and over time; we report rulebook ranges and archive all cited URLs with access dates to ensure verifiability.

hosting, and creator management are also required. Evidence from China, based on large samples, suggests that skill gaps are the primary obstacles. E-commerce retailers cite the lack of skills as one of the top three barriers, along with advertising costs and competition (World Bank, 2019). In LMICs, training and incubation are repeatedly identified as necessary complements to access to the digital market (Vidal and Faz, 2020).

Unlike commercial and social media channels, the GEP reduces operational complexity through a government-led model that bundles training, cooperative processing/packaging, and regional branding. In this setup, farmers are relieved of the burden of advertising, packaging design, and storefront competition, allowing them to focus on production while the program handles market-facing tasks. Instead, the local government centrally manages product promotion and brand development, marketing all agricultural products under the rural cooperative brand. In addition, the cooperative system allows farmers to convert their tea leaves into low-cost, market-ready standardized tea cakes, handling processing, packaging, and branding on their behalf. This integrated service framework substantially reduces skill demands for online sales; Farmers do not need to master performance marketing, live streaming, or complex digital operations.

References

World Bank (2019). *E-commerce Development: Experience from China* (Report No. 144689-CN). Washington, DC: World Bank Group. Available at https://documents1.worldbank.org/curated/en/552791574361533437/pdf/E-commerce-Development-Experience-from-China.pdf. Accessed: July 2025.

Vidal, M. and Faz, X. (2020). E-commerce is taking off in rural China: 3 lessons for other countries. *CGAP Blog*. Available at https://www.cgap.org/blog/E-commerce-is-taking-in-rural-china-3-lessons-for-other-countries. Accessed: July 2025.

Halaburda, H. and Yehezkel, Y. (2013). Platform competition under asymmetric information. *American Economic Journal: Microeconomics*, 5(3), 22–68. Available at https://www.aeaweb.org/articles?id=10.1257/mic.5.3.22. Accessed: July 2025.

Tmall Global (2024). Merchant onboarding and fee standards (security deposit, annual service fee, technical service fee rates). Official portal (in Chinese). Available at https://www.tmall.hk/wow/z/import/pegasus-no-head/S43HbztinhJ6JnTdYXW6. Accessed: July 2025.

JD Worldwide (2025a). POP transaction service fee (0.9%). Official Rule Center (in Chinese). Available at https://jdw-rule.jd.hk/detail?ruleId=950583665543483392. Accessed: July 2025.

JD Worldwide (2025b). Tiered security deposit management rules. Official Rule Center (in Chinese). Available at https://jdw-rule.jd.hk/detail?ruleId=950302479235551232. Accessed: July 2025.

Taobao (2024). Consumer Protection Service Agreement (includes deposit terms). Official terms (in Chinese). Available at https://terms.alicdn.com/legal-agreement/terms/suit_bu1_taobao/suit_bu1_taobao201709261344_28562.html. Accessed: July 2025.

Douyin E-commerce (2025a). Merchant technical service fee policy (category-based rates). Official Learning Center (in Chinese). Available at https://school.jinritemai.com/doudian/web/article/106833. Accessed: July 2025.

Douyin E-commerce (2025b). Affiliate (Jingxuan Alliance) settlement rules (general-plan commission 1–50%; targeted/shop-traffic plans up to 80%). Official Learning Center (in Chinese). Available at https://school.jinritemai.com/doudian/web/article/112620. Accessed: July 2025.

Appendix C: Timing of Adopting the GEP

Table C.1 provides a detailed overview of when different areas adopted the government-initiated e-commerce platform (GEP), identified by their respective area codes. The table lists the area codes (J1, J2, J3, J4, M1, M2) along with their respective platform access dates. For example, area J1 accessed the platform in June 2019, while area J2 did so in September 2018. Similarly, Area J3 accessed the platform in October 2020, and Area J4 in November 2019. The table also shows that Areas M1 and M2 accessed the platforms in November 2018 and April 2020, respectively. Overall, the table highlights the staggered pattern of the GEP access in different areas over time.

Table C.1. Timing of the GEP Access in Each Area

Area Code	**Platform Access Date**
J1	Jun 2019
J2	Sep 2018
J3	Oct 2020
J4	Nov 2019
M1	Nov 2018
M2	Apr 2020

Appendix D: Data Collection Process

The survey was carried out in two counties, designated as J and M, by a total of six specialized teams. The teams were selected by the local government and led by an area cadre or a local expert with expertise in tea farming. Their objective was to oversee the collection of data within a specified geographic area. The teams were composed of college and university students on vacation, as well as academically qualified local youth. A general manager was appointed to oversee the coordination of the survey and the subsequent consolidation of the data collected for each team.

Data were collected through in-person interviews, with each team member responsible for engaging with multiple households. The sample consisted of 983 households in the six selected areas, each of which received 20 RMB as an incentive to participate in the study. The survey encompassed a wide range of questions, including household characteristics, tea farming methods, marketing channels, and sources of household income. More than 90% of the respondents kept a household notebook to track relevant metrics, such as tea picking, farming output, and sales. Upon completing each household survey, the team leaders subjected the data to rigorous scrutiny to identify inconsistencies and ambiguities, which were then resolved before forwarding the collected information to the general manager for final aggregation.

Notes: The above images show our team's interactive sessions with local residents. The left photo captures our follow-up group gathering information at a local farmer's home. The right photo displays a notebook used by local farmers to record accounts, detailing the types of tea, sales volumes, and prices. After verification, team members log the data into our distributed forms based on different sales channels for various types of tea each year.

Figure D.1. Survey Engagement: Data Collection among Local Farmers

Figure D.1 provides a visual overview of the interview and data collection process conducted by the research team. The image shows a group of researchers and team members engaging in an interactive session with local farmers on the left side. Team members are seated around a table, participating in discussions and gathering information at what appears to be a local farmer's home. On the right side, a close-up shows a notebook used by local farmers, featuring handwritten records of various types of tea, sales volumes, and prices. According to the accompanying text, after verifying this information, the researchers input the data into distributed forms. These forms organize the information by different sales channels for various types of tea each year.

Before administering the survey, the team managers participated in comprehensive training sessions to ensure the integrity of the data. Following a comprehensive examination of the collected data, it was determined that the farming output and sales data, which represent more than 95% of the regional tea farming output, exhibited a high degree of alignment with the statistics reported in various media outlets. A comparison was made between the data obtained from the survey and publicly accessible news reports. The comparison is based on two core metrics: total tea output and its corresponding market value (that is, the level of agriculture output × price), covering the period 2016 to 2020. Regarding farming output, the mean yield in our dataset (964 tons) falls well within the range specified by news sources (870-1,480 tons). Similarly, the calculated average commercial value of the tea production output (495,733,250 RMB) is close to the values cited in the media reports (500 million RMB).[16]

[16] Sources for regional-level farming output and commercial values are as follows: https://www.chinanews.com.cn/cul/2014/08-26/6529253.shtml (accessed on 27 August 2023); http://www.puernews.com/zthd/pejmsgcysw/0311009048285368 8837 (accessed on 27 August 2023); https://m.puercn.com/show-8-44415.html (accessed on 27 August 2023).

Appendix E: Household- and Area-related Statistics

Table E.1 summarizes the household and area-level variables in our data. The results show that plot sizes, as well as local infrastructure, did not change significantly before and after 2018. This suggests that local market conditions remained relatively stable during our sample period, aside from the introduction of the GEP and its associated public services.

Table E.1. Summary Statistics for Household- and Area-level Variables

	Before 2018	**After 2018**
Acres of Tea Trees	17.00	16.87
	(7.76)	(7.77)
Acres of Tea Gardens	34.70	34.56
	(11.34)	(11.31)
Operating Factories	14.12	15.71
	(7.30)	(7.72)
Shipping Companies	3.22	4.07
	(1.26)	(1.28)

Notes: We report the standard deviation in parentheses.

Appendix F: Robustness Check 1: Unobserved Trends and Environmental Changes

In this section, we extend our baseline specification by adding household and area-level controls as well as county-specific time trends. The results of the estimation are shown in Table F.1. The estimated treatment effects, after including the additional controls, align with the results of the baseline specification. Specifically, the data show that online sales increase by an average of 18.4122% after gaining access to the GEP. In contrast, offline sales decrease by an average of 16.222% after access to the platform. In Column (3), we control for household-level farming output (volume), while in Column (4), we account for both volume and area characteristics, such as the number of factories and shipping companies. These findings are consistent with our previous results, indicating that the increase in online sales can be largely attributed to the GEP rather than changes in production technology or the local market.

Table F.1. Effect of the GEP Access on Sales with Additional Controls

Dependent Variable:	Log(sales): $q_{i,j,t}$			
	Time-varying Controls		**County-specific Trends**	
	(1)	(2)	(3)	(4)
Online Sales	-0.522***	-0.513***	-0.522***	-0.513***
	(0.035)	(0.061)	(0.035)	(0.061)
Platform Access	-0.183***	-0.177***	-0.183***	-0.181***
	(0.038)	(0.047)	(0.036)	(0.048)
Platform Access × Online Sales	0.356***	0.346***	0.356***	0.346***
	(0.084)	(0.099)	(0.084)	(0.099)
Zero Output	-5.153***	-5.154***	-5.153***	-5.154***
	(0.087)	(0.087)	(0.087)	(0.087)
Log(Volume)	0.052***	0.052***	0.052***	0.052***
	(0.005)	(0.005)	(0.005)	(0.005)
Number of Operating Factories		0.002		0.013
		(0.004)		(0.008)
Number of Shipping Companies		0.006		-0.005
		(0.014)		(0.020)
Number of Factories × Online Sales		-0.003		-0.003
		(0.004)		(0.004)
Number of Companies × Online Sales		0.010		-0.010
		(0.019)		(0.019)
Observations	29,490	29,490	29,490	29,490
Quality FE	YES	YES	YES	YES
Household FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
County Specific Trend	NO	NO	YES	YES
R^2	0.966	0.966	0.966	0.966

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. Significance levels are denoted as follows: ∗p < 0.10, ∗∗p < 0.05, ∗∗∗p < 0.01.

Appendix G: Robustness Check 2: Treatment Endogeneity

As shown in Table G.1, the results indicate that the proposed model explains approximately 76% of the observed variation in the adoption of the platform. Next, area-specific, time-varying factors that can be linked to the timing of platform adoption are added, including the total tea production, the number of factories, and the number of shipping companies. No statistically significant coefficients were found for these factors, and their inclusion did not improve the explanatory power of the regression, suggesting that the timing of treatment is not related to area-specific and time-varying factors.

Table G.1. Likelihood of the GEP Access

Dependent Variable:	Access to the GEP	
	(1)	(2)
2018	0.333*	0.241
	(0.149)	(0.167)
2019	0.667***	0.571***
	(0.149)	(0.169)
2020	1.000***	0.876***
	(0.149)	(0.177)
Volume of Tea Produced		-0.033
		(0.070)
Number of Factories		-0.005
		(0.007)
Number of Shipping Companies		0.057
		(0.036)
Observations	48	48
R^2	0.763	0.779

Notes: Standard errors are indicated in parentheses. This table reports the estimated coefficients when regressing treatment status (access to the platform) on year-fixed effects and area-level characteristics. Including area-specific characteristics does not increase the explanatory power of the model once we control for year effects. Significance levels are denoted as follows: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

We also present the results of the placebo tests to better interpret the treatment. In our first placebo test, we randomized the years during which a household or area had access to the platform, while keeping the total number of years of access unchanged. These results are shown in Columns (1) and (2) of Table G.2. In Column (1), we shuffle treatment at the area level. For example, if an area had access to the GEP in 2019 and 2020 (a two-year period), we randomly select two years between 2016 and 2020 and assign a value of one to a new variable called "placebo treatment" for those years. The placebo treatment is applied uniformly to all households in that area. In Column (2), treatment status is reshuffled for each household instead of each area. After creating the placebo treatment, we estimate its effect on offline and online sales. Both columns indicate that the placebo treatment has no statistically significant impact on online or offline sales of households at the 10% significance level. In the second placebo test, we estimate Equation 1 in our manuscript using a subset of households that have never participated in online sales during the entire sample period. Our data indicate that about 9% of the total sample falls into this group. If the impact of the GEP on tea sales across different channels is solely due to the introduction of the platform, these non-online sellers should remain unaffected by the policy change.

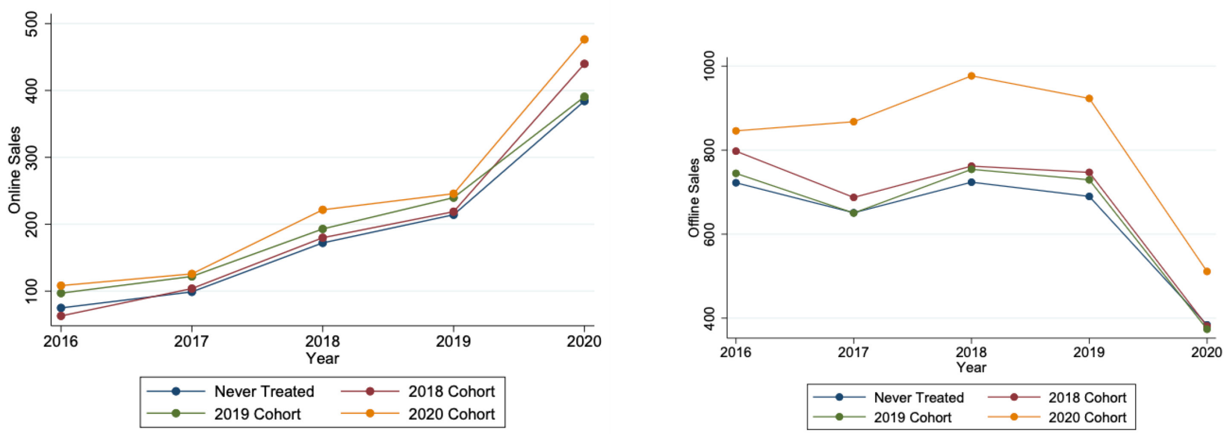
Table G.2. Placebo Tests: Effect of Placebo Treatment on Sales

Dependent Variable:	Log(sales): $q_{i,j,t}$		
	Re-shuffled Treatment		**Non-adopters**
	Area Level	Household Level	
	(1)	(2)	(3)
Platform Access	-0.067	-0.002	-0.017
	(0.084)	(0.020)	(0.013)
Platform Access × Online Sales	0.123	0.014	
	(0.162)	(0.029)	
Online Sales	-0.426***	-0.396***	
	(0.064)	(0.021)	
Zero Output	-5.438***	-5.439***	-4.689***
	(0.068)	(0.069)	(0.108)
Observations	29,490	29,490	29,490
Household FE	YES	YES	YES
Quality FE	YES	YES	YES
Year FE	YES	YES	YES
R^2	0.964	0.965	0.948

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. Significance levels are denoted as follows: ∗p < 0.10, ∗∗p < 0.05, ∗∗∗p < 0.01.

Appendix H: Robustness Check 3: Parallel Trends

To further ensure that our estimated effects are causal, we demonstrate that online and offline sales across different areas would have followed similar patterns (parallel trends) in the absence of the GEP. First, we plot the evolution of online and offline sales based on the year they first gained access to the platform (cohorts). As shown in Figure H.1, both online and offline sales exhibit similar trends in the pretreatment periods, with online sales increasing and offline sales decreasing during this time. This indicates that without the GEP, online sales would have increased at roughly similar rates across different areas, and any additional growth in online sales beyond this is attributable to the introduction of the platform.



Notes: The above figure plots the evolution of online and offline sales for different cohorts.

Figure H.1. Offline and Online Sales Trends

We further verify this result by estimating the marginal effects of time (trend) on online sales across cohorts and then testing whether these estimated trends differ among cohorts. These estimated trends are shown in Table H.1. Using a Wald test, we fail to reject the null hypothesis that these pre-trends are equal, providing additional evidence that the evolution of online sales is consistent across cohorts before the launch of the GEP.

Table H.1. Estimated Pre-trends by Cohort

Dependent Variable:	Cohort Mean Online Sales: $\bar{q}_{c,online,t}$			
	Never Treated	2018 Cohort	2019 Cohort	2020 Cohort
	(1)	(2)	(3)	(4)
Marginal Effect of t	48.623***	58.433***	47.985***	56.653***
	(15.753)	(15.753)	(15.753)	(15.753)

Notes: Standard errors are indicated in parentheses. The dependent variable is the average online sales in each cohort over time. Significance levels are denoted as follows: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

Appendix I: Robustness Check 4: Bias Correction Related to TWFE Estimators

I.1 Negative Treatment Weights

Our analysis examines the staggered adoption of the platform across different villages. To control for household-specific, year-specific, and quality-specific shocks, we include fixed effects. However, literature such as De Chaisemartin and d'Haultfoeuille (2020) and Jakiela (2021) warns of potential bias in treatment effect estimates when effects vary over time or between units. In this section, following Jakiela (2021), we show that our treatment effect estimates remain unbiased after including household-, quality-, and year-fixed effects.

We base our analysis on the following equation:

$$q_{i,j,t} = \alpha + \gamma D_{i,t} + \delta mode_{i,j,t} + \theta D_{i,t} \times mode_{i,j,t} + \zeta Z_{i,j,t} + \mu_i + \eta_j + \psi_t + \epsilon_{i,j,t}, \tag{I.1}$$

where $\hat{\theta}^{TWFE}$, the OLS estimator for treatment effect θ, can be derived using the Frisch-Waugh-Lovell theorem:

$$\hat{\theta}^{TWFE} = \sum_{\mathrm{i,j,t}} q_{i,j,t} \left(\frac{\hat{\epsilon}_{i,j,t}}{\sum_{i,j,t} \hat{\epsilon}^2_{i,j,t}} \right), \tag{I.2}$$

where $\hat{\epsilon}_{i,j,t}$ representing the residual from regressing the treatment indicator on the household-, year-, and quality-fixed effects. The treatment effect is therefore a weighted sum of the outcome variable, with the weights being the residualized treatment weights. Jakiela (2021) states that bias occurs when treated units have negative treatment weights and when treatment effects vary.

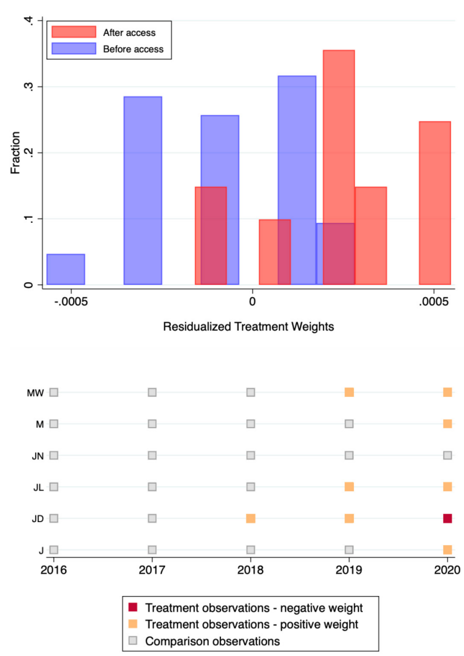


Figure I.1. Weights of Two-Way Fixed Effects

To detect such biases, we check whether treated units have negative weights and then test for homogeneity of treatment effects. Following Jakiela (2021), we regress our treatment indicator on the fixed effects to obtain the residualized treatment $\sum_{i,j,t} \hat{\epsilon}_{i,j,t}^2$. We then construct the treatment weights $\sum_{i,j,t} \hat{\epsilon}_{i,j,t}^2$ for each observation. Figure I.1 displays these weights for treated and untreated units. Figure I.1 shows these weights for treated and untreated units. The figure indicates that only 15% of the treated units have negative weights. For context, Jakiela (2021) found that about 25% of the treated units had negative weights, yet the treatment effect remained strong after removing these observations. Since our Average Treatment Effect (ATE) estimate is a weighted sum of outcomes, these small negative weights are unlikely to cause bias.

Table I.1. Effect of the GEP access on Sales (Negative Treatment Weights Excluded)

Dependent Variable:	Log(sales): $q_{i,j,t}$		
	(1)	(2)	(3)
Online Sales	-0.474***	-0.474***	-0.482***
	(0.036)	(0.036)	(0.037)
Platform Access	-0.127*	-0.160*	-0.150**
	(0.061)	(0.073)	(0.050)
Platform Access × Online Sales	0.282**	0.282**	0.288**
	(0.109)	(0.109)	(0.111)
Zero Output	-5.482***	-5.482***	-5.429***
	(0.077)	(0.077)	(0.068)
Constant	5.738***	5.746***	5.715***
	(0.096)	(0.098)	(0.057)
Observations	28,284	28,284	28,284
Quality FE	NO	NO	YES
Household FE	NO	NO	YES
Year FE	NO	YES	YES
R^2	0.955	0.956	0.965

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. Significance levels are denoted as follows: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

As a further robustness check, we recalculated our model excluding treated units with negative weights. The revised results in Table I.1 confirm a significant substitution effect after the launch of the platform: offline sales decreased by approximately 13.929%, and online sales increased by approximately 14.798%.

I.2 Interaction Weighted Estimator

To further address potential bias in two-way fixed effects estimators, we also used the interaction-weighted (IW) fixed effects estimator, as suggested by Sun and Abraham (2021) and Callaway and Sant'Anna (2021). This estimator is robust to varying treatment effects in models with staggered treatment timing and can be applied even when there is no never-treated group. Following the approach of Sun and Abraham (2021), we divided our sample into distinct cohorts based on the year each household started using the platform. In our study, this creates three cohorts (2018, 2019, and 2020) plus a group that was never treated. We first estimate the effect of the average treatment effect over time in the treated units (CATT) using a two-way fixed effects model that interacts with cohort indicators with a relative period indicator. These relative period indicators show how many periods each cohort has been treated, allowing treatment effects to change over time. For a static model, an alternative estimate of CATT can be used, where cohort indicators interact with a binary treatment indicator.

The following equation is estimated:

$$\begin{aligned} q_{i,j,t} = \alpha + \sum_{e \notin C} \sum_{l=-1}^{2} \gamma_{e,l} \left(1\{E_i = e\} \cdot D_{i,t}^{l}\right) + \delta mode_{i,j,t} + \\ \textstyle\sum_{e \notin C} \sum_{l=-1}^{2} \theta_{e,l} \left(1\{E_i = e\} \cdot D_{i,t}^{l}\right) + mode_{i,j,t} + \\ \zeta Z_{i,j,t} + \beta' X_{i,t} + \mu_i + \eta_j + \psi_t + \epsilon_{i,j,t}, \end{aligned} \tag{I.3}$$

where $E_i \in \{2018, 2019, 2020, \infty\}$ denotes the year that household i first gained access to the platform (treatment), C is the set of households that were never treated and $D_{i,t}^{l}$ is an indicator for household i being l periods away from treatment in period t.

Subsequently, the weights were calculated based on the sample share of each cohort in each relative period. Ultimately, the IW estimate of the treatment effect is derived by weighting the average of the CATT using the weights obtained in the previous step. The IW estimates are shown in Table I.2. The results of our analysis, which uses the IW two-way fixed effects estimator, suggest that the impact of the GEP on tea sales aligns with our baseline findings. Specifically, the estimated coefficient for platform access is -0.156, while the estimate for the interaction between platform access and online sales is 0.274. These coefficients were converted into effects on online and offline sales, resulting in a 14.444% decrease in offline sales and a 12.524% increase in online sales. Both estimated treatment effects are statistically and economically significant, supporting the hypothesis that farmers shifted their sales from offline to online channels after gaining access to the GEP.

Table I.2. Interaction Weighted TWFE Estimates

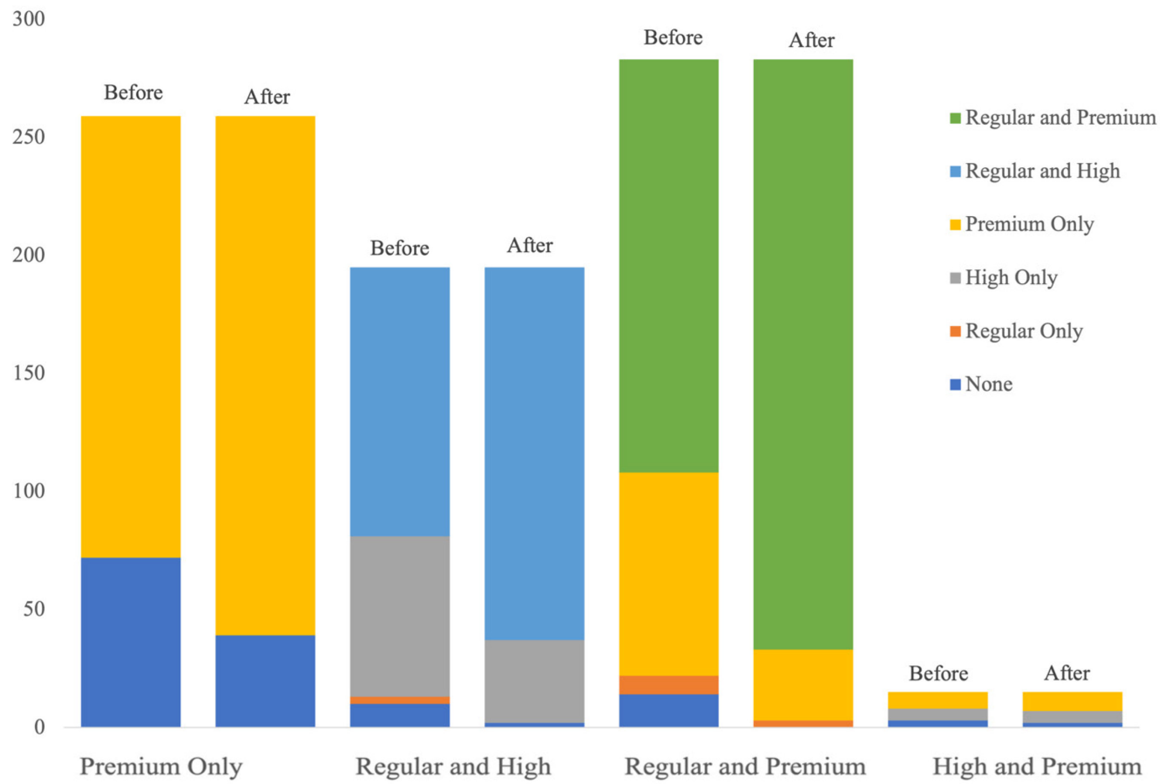
Dependent Variable:	Log(sales): $q_{i,j,t}$		
	(1)	(2)	(3)
Online Sales	-0.474***	-0.474***	-0.482***
	(0.043)	(0.043)	(0.045)
Platform Access (γ)			
Cohort 1, $t_0 - 1$	-0.104	-0.056	-0.026
	(0.060)	(0.070)	(0.046)
Cohort 1, t_0	-0.062	-0.107	-0.070
	(0.060)	(0.063)	(0.039)
Cohort 1, $t_0 + 1$	-0.083	-0.044	-0.097**
	(0.060)	(0.024)	(0.036)
Cohort 1, $t_0 + 2$	-0.263***	-0.224***	-0.281***
	(0.053)	(0.019)	(0.032)

Cohort 2, $t_0 - 1$	-0.041	-0.086	-0.034
	(0.109)	(0.118)	(0.052)
Cohort 2, t_0	-0.056	-0.017	-0.054
	(0.112)	(0.106)	(0.053)
Cohort 2, $t_0 + 1$	-0.273**	-0.234**	-0.275***
	(0.092)	(0.085)	(0.034)
Cohort 3, $t_0 - 1$	0.079	0.118***	-0.003
	(0.051)	(0.024)	(0.020)
Cohort 3, t_0	-0.145**	-0.106***	-0.232***
	(0.047)	(0.016)	(0.020)
Interaction Weighted	-0.129***	-0.098***	-0.156***
	(0.053)	(0.028)	(0.028)
Platform Access × Online Sales (θ)			
Cohort 1, $t_0 - 1$	0.007	0.007	0.007
	(0.054)	(0.054)	(0.054)
Cohort 1, t_0	0.046	0.046	0.048
	(0.051)	(0.051)	(0.052)
Cohort 1, $t_0 + 1$	0.153**	0.153**	0.155**
	(0.051)	(0.051)	(0.052)
Cohort 1, $t_0 + 2$	0.515***	0.515***	0.525***
	(0.041)	(0.041)	(0.043)
Cohort 2, $t_0 - 1$	0.036	0.036	0.039
	(0.079)	(0.079)	(0.079)
Cohort 2, t_0	0.130	0.130	0.133
	(0.072)	(0.072)	(0.072)
Cohort 2, $t_0 + 1$	0.463***	0.463***	0.473***
	(0.051)	(0.051)	(0.054)
Cohort 3, $t_0 - 1$	-0.019	-0.019	-0.016
	(0.039)	(0.039)	(0.040)
Cohort 3, t_0	0.425***	0.424***	0.435***
	(0.039)	(0.039)	(0.042)
Interaction Weighted	0.268***	0.268***	0.274***
	(0.042)	(0.042)	(0.044)
Zero Output	-5.479***	-5.480***	-5.424***
	(0.073)	(0.073)	(0.066)
Constant	5.733***	5.722***	5.718***
	(0.097)	(0.083)	(0.061)
Observations	29,490	29,490	29,490
Quality FE	NO	NO	YES
Household FE	NO	NO	YES
Year FE	NO	YES	YES
R^2	0.956	0.956	0.965

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. The areas are divided into cohorts based on the year in which they were treated. For this study, the term "treatment" is defined as having access to the platform for a minimum of four consecutive calendar months within a given year. Area J2 (as of 09.2018) is included in Cohort 1. Areas M1 (as of 11.2018) and J1 (as of 06.2019) are included in Cohort 2. Areas J4 (as of 11.2019) and M2 (as of 04.2020) are included in Cohort 3. Area J3 (as of 10.2020) is not included in the study and serves as a control group. Significance levels are denoted as follows: *p < 0.10, **p < 0.05, ***p < 0.01.

Appendix J: Additional Evidence on Product Mix

Figure J.1 summarizes the model-free household counts based on the bundles of quality tea sold by farmers. It also shows, within each bundle, the set of qualities they sell online before (before 2018) and after gaining access to the GEP (2018 onward). The distribution reflects a shift toward listing lower-priced products online once the public e-commerce service becomes available.



Notes: Bars display the number of households by four production bundles: (1) premium only, (2) regular and high, (3) regular and premium, and (4) high and premium. Within each bundle, colors indicate which qualities are available for sale online. "Before" refers to pre-2018, and "After" indicates 2018 onward. Online sales include transactions through the GEP, commercial platforms, and social media channels. For households producing both regular and premium tea, the number of households selling regular tea online increases after gaining access, while the number of households selling only premium tea online decreases.

Figure J.1. Qualities Sold Online Before and After Launching the Platform

The shift is most pronounced among households that produce regular and premium tea. Before 2018, of the 283 such households, 175 sold both types online, 86 sold only premium-quality tea, 8 sold only regular tea, and 14 sold neither. After 2018, 250 sold both types online, 30 sold only premium-quality tea, and 3 sold only regular tea (none remained offline). The decline in premium-only and offline-only segments, together with the growth of the "both channels" segment, is consistent with the GEP lowering the costs of selling lower-priced regular tea online (e.g., through cooperative packaging and public branding). This compositional shift aligns with our regression results, which show higher online sales for both regular and premium teas.

Appendix K: Effects Across Different Pretreatment Channel of Sales

To gain further insight into the role of the GEP, we examine its effect on farmers by dividing them based on the markets where they sold their tea before the GEP was introduced. Farmers are initially divided into two groups based on their online sales channels prior to treatment. The first group consists of farmers who only sold tea online through social media. The second group includes farmers who used commercial platforms for tea sales before gaining access to the GEP.

A large portion of farmers in the second group also sell their tea through social media platforms. However, we observe that most farmers who sell on social media do not use formal e-commerce platforms. We believe this is because commercial e-commerce platforms often set entry barriers to filter out high-quality merchants. These barriers effectively prevent farmers in rural areas from selling low-end products online. Therefore, we hypothesize that the barriers to online sales are lower for farmers selling high-quality or premium-quality tea compared to selling regular tea on online platforms.

Table K.1. Heterogeneous Effects of the GEP Access on Sales by Pretreatment Online Channels and Quality

Dependent Variable:	Regular Tea Sales		High-quality Tea Sales		Premium-quality Tea Sales	
	Social Media	**Platform**	**Social Media**	**Platform**	**Social Media**	**Platform**
	(1)	(2)	(3)	(4)	(5)	(6)
Platform Access	−0.103***	−0.236**	−0.171***	−0.246**	−0.193**	−0.418
	(0.012)	(0.064)	(0.039)	(0.074)	(0.048)	(0.310)
Platform Access × Online Sales	0.208***	0.314	0.316**	0.435**	0.412***	0.542*
	(0.030)	(0.164)	(0.080)	(0.130)	(0.075)	(0.249)
Online Sales	−0.589***	−0.385**	−0.509***	−0.477**	−0.719***	−0.406**
	(0.037)	(0.093)	(0.085)	(0.167)	(0.123)	(0.108)
Zero Output	−5.868***	−6.254***	−5.067***	−5.343***	−5.073***	−5.418***
	(0.058)	(0.032)	(0.095)	(0.093)	(0.081)	(0.148)
Observations	7,510	610	7,510	610	7,510	610
Household FE	YES	YES	YES	YES	YES	YES
Quality FE	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES
R^2	0.974	0.982	0.978	0.980	0.974	0.976

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. Significance levels are denoted as follows: ∗p < 0.10, ∗∗p < 0.05, ∗∗∗p < 0.01.

Table K.1 illustrates the impact of access to GEP on tea quality, including regular, high-quality, and premium-quality, among farmers who used only social media for sales, compared to those who used commercial platforms before the introduction of GEP. In Columns (1) and (2), the results indicate that the increase in online sales of regular tea is statistically significant for farmers who previously sold through social media. However, this significance does not apply to farmers who use commercial platforms. In contrast, Columns (3)-(6) reveal that the increase in online sales of high- and premium-quality tea is statistically significant (at the 10% level) for both groups.

Overall, the table shows that the only exception is the subgroup that was already qualified to sell on commercial marketplaces before the program: for these farmers, GEP access has minimal additional impact. They tend to operate on a larger scale, possess the necessary qualifications, and are almost always active on social media; thus, the public storefront does not significantly boost their margins or profits. These findings support our previous findings: the GEP provides a low-cost alternative channel for farmers who previously could not profitably sell lower-end teas on commercial platforms, allowing them to sell their products online.

Appendix L: Reconciling the Effect of GEP access

This section clarifies why, in our manuscript, the effect of the GEP access on online sales reported in Table 2 (0.166) exceeds the corresponding estimate in Table 8 (0.081). Both tables capture the same underlying effect, but they do so under different aggregation and functional-form choices. Once these differences are made explicit, the numerical gap is expected.

L.1 Reconciling the Log Effects

We index households by i, years by t, and product qualities by j. Let $q_{i,j,t} \geq 0$ denote household i's sales of quality j in year t. We define $\text{Online}_{i,j,t} \in \{0,1\}$ as a binary online channel indicator and $\text{Platform}_{i,t} \in \{0,1\}$ as a binary indicator for access to the GEP. We write the total online sales volume as

$$Z_{i,t} \equiv \sum_j q_{i,j,t}^{\text{online}} = \sum_j q_{i,j,t} \cdot \text{Online}_{i,j,t}.$$

At the quality level (Table 2), we estimate Equation L.1, a "log of parts" specification in which the effect of access to the GEP on online sales is $\beta \equiv b_0 + c_0$. The estimating equation is

$$\log\left(q_{i,j,t} + 1\right) = \alpha + b_0 \text{Platform}_{i,t} + c_0\left(\text{Platform}_{i,t} \times \text{Online}_{i,j,t}\right) + d_0 \text{Online}_{i,j,t} + \text{FE}s + \varepsilon_{i,j,t}. \quad \text{(L.1)}$$

The marginal effect of access to the platform on online sales in Equation L.1 is $\beta \equiv b_0 + c_0$.

At the household-year aggregate, we estimate Equation L.2, a "log of sum" specification where b_1 captures the effect of access to the GEP on $\log\left(1 + Z_{i,t}\right)$. The estimating equation is

$$\log\left(Z_{i,t} + 1\right) = \alpha_1 + b_1 \text{Platform}_{i,t} + \text{FE}s + u_{i,t}, \quad \text{(L.2)}$$

so b_1 is the effect of the access to the platform on $\log\left(1 + Z_{i,t}\right)$.

Equations L.1 and L.2 are not algebraically equivalent because they apply the concave link $\log(1 + \cdot)$ to different objects: $\log(1 + \text{part})$ versus $\log(1 + \sum \text{parts})$. Concavity implies mechanical compression when moving from the former to the latter. To see this, suppose that the platform scales the sales of each quality online by the same factor e^{β} (with $\beta = b_0 + c_0$): $q_{i,j,t}^{\text{online}} \to e^{\beta} q_{i,j,t}^{\text{online}}$. Let $Z \equiv Z_{i,t}$ denote the total volume of online sales before access to the platform. The induced change in the aggregate dependent variable is

$$\Delta(\beta; Z) = \log\left(1 + e^{\beta} Z\right) - \log(1 + Z) = \log\left(1 + \left(e^{\beta} - 1\right)\frac{Z}{1 + Z}\right). \quad \text{(L.3)}$$

Because $Z/(1+Z) \in (0,1)$, we have $\Delta(\beta; Z) < \beta$ for every $Z > 0$, with strict inequality unless $Z \to \infty$. Thus, even if the quality-level log effect equals β, the aggregate "log of sum" effects is strictly smaller. For small to moderate β, a first-order expansion of Equation L.3 yields

$$\Delta(\beta; Z) \approx \beta \cdot \frac{Z}{1+Z}. \tag{L.4}$$

Taking expectations on the household-year distribution of Z, we have

$$\mathrm{b}_1 \approx \beta \cdot \mathbb{E}[\frac{Z}{1+Z}]. \tag{L.5}$$

L.2 Quantitative Approximation with Zero Observations

Table 8 includes that 2,266 out of 4,915 household-year aggregates have zero online sales. Under the approximation of Equation L.4, the mapping between the log effect at the quality level β and the aggregate log effect $b_1 \approx \beta \cdot \mathbb{E}[Z/(1+Z)]$. If, as the data suggest, $Z/(1+Z)$ is essentially 0 when $Z = 0$ and is close to 1 for most non-zero observations, then

$$\mathbb{E}\left[\frac{Z}{1+Z}\right] \leq (1-p_0) \cdot 1 + p_0 \cdot 0 = 1 - \frac{2266}{4915} = 0.539.$$

Using the new estimates $\beta \approx 0.166$ and $b_1 \approx 0.081$,

$$\frac{b_1}{\beta} = \frac{0.081}{0.166} = 0.488,$$

which is very close to $\mathbb{E}[Z/(1+Z)] \approx 0.5$. Hence, the attenuation from the quality–level estimate to the aggregate estimate is quantitatively explained by the mass of zeros and the concavity $\log(1+\cdot)$ embodied in Equation L.4.

Appendix M: Additional Mediation Analysis

This appendix complements Section 6.4 in the chapter 1 by exploring two mechanism questions. First, we examine whether changes in local logistics—measured by the presence of shipping companies—mediate the program's effect on online sales. If local logistics are also a key pathway, including shipping company counts should reduce the GEP coefficient and improve model fit, even before adding online channel breadth or online product variety. Second, we verify whether the mediation pattern is symmetric when the outcome shifts from total online sales to total offline sales: If the local GEP shifts transactions across channels, the same mediators should explain the offline declines.

M.1 Mediating Role of Shipping Companies on Online Sales

Table M.1 reports household–year regressions with the log of total online sales as the dependent variable. All columns include household- and year-fixed effects, as well as an indicator for zero output; standard errors are clustered at the area level. Column (1) presents the baseline specification. Column (2) adds the count of shipping companies to capture contemporaneous changes in local logistics capacity. Columns (3) and (4) separately add the two hypothesized mediators—number of online channels and number of online product quality varieties—while column (5) includes both mediators jointly alongside shipping companies.

Table M.1. Mediating Role of Shipping Companies

Dependent Variable:	Total Online Sales				
	(1)	(2)	(3)	(4)	(5)
Platform Access	0.081*	0.095*	0.029	0.037	0.017
	(0.035)	(0.046)	(0.062)	(0.030)	(0.039)
Shipping Companies		0.084	-0.019	0.020	-0.018
		(0.078)	(0.066)	(0.013)	(0.031)
Number of Channels			2.358***		1.218***
			(0.085)		(0.055)
Number of Varieties				2.578***	1.971***
				(0.062)	(0.035)
Zero Output	-2.387***	-2.388***	-0.952***	-1.031***	-0.609***
	(0.191)	(0.188)	(0.106)	(0.102)	(0.065)
Observations	4,915	4,915	4,915	4,915	4,915
Household FE	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES
R^2	0.777	0.777	0.884	0.926	0.946

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. Significance levels are denoted as follows: *p < 0.10, **p < 0.05, ***p < 0.01.

The evidence does not support shipping companies as a mediator. Adding shipping companies in column (2) leaves the GEP coefficient essentially unchanged, and the shipping company coefficient is small and statistically insignificant. The model fit also remains the same, indicating that logistics alone does not add explanatory power. In contrast, when we add the number of online channels in Column (3) or online varieties in Column (4), each mediator enters strongly and positively. The model R-squared coefficient increases significantly (to 0.884 and 0.926), and the GEP coefficient decreases toward zero (0.029 and 0.037, both statistically indistinguishable from zero). In Column (5), both mediators remain highly significant, while the shipping-company coefficient remains small and insignificant, and the GEP coefficient further decreases to 0.017. Overall, these patterns suggest that the expansion of online channels and the increase in online varieties, rather than changes in the presence of local shipping companies, are the primary pathways through which the program boosts online sales.

M.2 Mediating Role of Online Channels and Product Variety on Offline Sales

Table M.2 repeats the mediation design with the log of total offline sales as the dependent variable. Column (1) reports the baseline; Column (2) adds shipping companies; Columns (3) and (4) separately add the number of online channels and the number of online varieties; Column (5) includes both mediators jointly.

Table M.2. Mediating Role of Online Channels and Product Variety on Offline Sales

Dependent Variable:	Total Offline Sales				
	(1)	(2)	(3)	(4)	(5)
Platform Access	-0.061*	-0.062*	-0.051	-0.059*	-0.052
	(0.026)	(0.029)	(0.030)	(0.029)	(0.030)
Shipping Companies		-0.006	0.010	-0.003	0.010
		(0.036)	(0.037)	(0.033)	(0.038)
Number of Channels			-0.364***		-0.409***
			(0.064)		(0.079)
Number of Varieties				-0.126***	0.078*
				(0.027)	(0.037)
Zero Output	-4.895***	-4.895***	-5.117***	-4.962***	-5.104***
	(0.115)	(0.115)	(0.111)	(0.120)	(0.116)
Observations	4,915	4,915	4,915	4,915	4,915
Household FE	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES
R^2	0.833	0.833	0.837	0.833	0.837

Notes: Standard errors are indicated in parentheses. Error terms are clustered at the area level. Significance levels are denoted as follows: ∗p < 0.10, ∗∗p < 0.05, ∗∗∗p < 0.01.

The results mirror the online analysis in two ways. First, shipping companies do not influence the effect: including them in Column (2) leaves the GEP coefficient basically unchanged, and the shipping coefficient is small and not significant, with no improvement in model fit. Second, online mediators predict offline declines. When added separately, more online channels and greater online variety are each linked to lower offline sales, and adding either reduces the size and significance of the GEP coefficient. In the combined model, channels stay negative and highly significant, while the variety coefficient becomes small and positive. Including both mediators renders the GEP coefficient statistically indistinguishable from zero, indicating that the offline drop is primarily explained by the combined growth in online channels and product variety.

Appendix N: Price Imputation Strategy

Our dataset does not contain renewal prices for users who chose not to renew their subscriptions. This section explains the methodology we used to impute these unobservable prices, balancing precision and comprehensiveness. We have two primary goals in our price imputation approach. Firstly, we aim for precision to accurately reflect the prices users would have faced during their decision-making period. Secondly, we strive for comprehensive coverage, aiming to match prices for a wide range of users, particularly since our dataset includes detailed information on user referral networks. This is crucial to avoid attrition that might introduce bias in estimating network effects. Accordingly, we adopt time windows of varying widths as our price imputation strategy.

Figure N.1. Community-Level Price Adjustment Frequencies

For each user in a community without an observable renewal price, we compute the average renewal prices for users within that community who renewed within $\pm w$ days of their subscription expiration. Recognizing the potential for inaccuracies in average-based imputation due to price volatility, our analysis found that most communities adjust their prices about once a month or more often. This is illustrated in Figure N.1, where a representative community's average price change frequency is once every 8.84 months. Therefore, we believe that imputed prices based on these averages are a reasonable approximation. Considering the platform's provision allowing renewals up to 14 days prior to expiration, we examine price imputation under various time windows, specifically, $w = 3,7,14,30$ days.

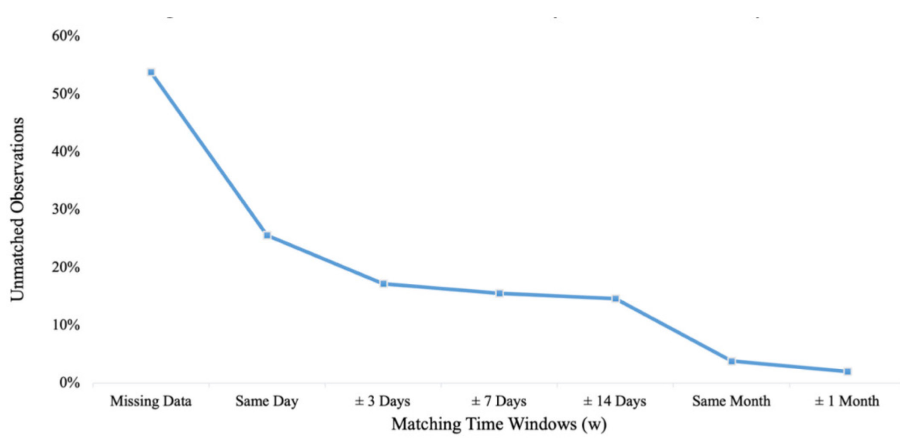


Figure N.2. Trade-off Between Accuracy and Data Recovery

Figure N.2 shows how the quantity of missing price data captured varies with different time windows (w). This figure highlights the trade-off between accuracy and data coverage. Using the renewal price on the exact day of a user's subscription expiration would be the most precise but limits data recovery to about 30% of total observations. As we expand the time window, the proportion of matchable data increases. Notably, a 14-day window serves as a critical point where data coverage significantly improves.

Table N.1. OLS Estimation Results across Different Matching Windows

Dependent Variable:	Renewal Decision			
	(1)	(2)	(3)	(4)
Price	-0.066***	-0.068***	-0.058***	-0.066***
	(0.003)	(0.003)	(0.003)	(0.003)
Referrer's Decision	0.043***	0.045***	0.030***	0.027***
	(0.009)	(0.009)	(0.007)	(0.009)
Referred	-0.034***	-0.033***	-0.050***	-0.033***
	(0.004)	(0.004)	(0.007)	(0.004)
Number of Referees	0.127***	0.127***	0.149***	0.140***
	(0.004)	(0.004)	(0.005)	(0.004)
Observations	255,244	260,378	263,161	302,127
Joining Time FE	YES	YES	YES	YES
Renewal Time FE	YES	YES	YES	YES
Community FE	YES	YES	YES	YES
R^2	0.316	0.323	0.323	0.296
Matching Window	$\pm$3 Days	$\pm$7 Days	$\pm$14 Days	$\pm$1 Month
$\pm$3 Days	√	√	√	√
$\pm$7 Days		√	√	√
$\pm$14 Days			√	√
$\pm$1 Month				√
Unmatched Observations	17.20%	15.54%	14.46%	1.99%
Unmatched Referrer Percentage	1.14%	0.86%	0.73%	0.36%
Unmatched Referee Percentage	6.68%	6.05%	5.75%	1.62%

Notes: Standard errors are indicated in parentheses. Significance levels are denoted as follows: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

We conducted additional studies to determine the most suitable time window for price imputation. Regression models were constructed for each time window, including key variables such as user renewal decisions, price, referrer decisions, referral status, and the count of referees. We also accounted for various fixed effects like community-level and time-level fixed effects to enhance model robustness. Our criterion for the ideal price-imputation strategy aims to preserve the price accuracy while maximizing the scope of matched renewal prices and observations. Table N.1 presents the variations in regression coefficients, matched and unmatched observations, and the subset of unmatched observations with existing referral relationships for different values of w. A 14-day window, representing the average monthly price, yielded statistically robust coefficients and the highest R^2 values without compromising data volume. While extending w to a month increased the number of matched observations, the reduced precision in price led to a decrease in R^2 values. Consequently, we chose a 14-day window as our standard for data construction and empirical analyses.

Appendix O: Multi-Homing Analysis

This section investigates the extent to which users on our platform engage in multi-homing, that is, subscribing to multiple communities within the same period. This behavior could introduce strategic interdependencies across different communities during decision-making. Given the similarity of our platform to emerging content-monetization platforms like Patreon, which rely heavily on private traffic generated by influencers (community owners), understanding multi-homing is crucial. Conversations with platform managers have indicated that multi-homing behavior is relatively uncommon among users.

To empirically assess this claim, we conducted a thorough analysis of our dataset. Our findings align with the managerial insights, confirming that multi-homing is indeed infrequent. Specifically, we found that only 2.21% of users are subscribed to more than one community in the same month over the observed year. Within this group, 91.78% are members of no more than two communities, accounting for a mere 2.02% of the total user base. Figure O.1 illustrates the distribution of multi-homed users across the platform.

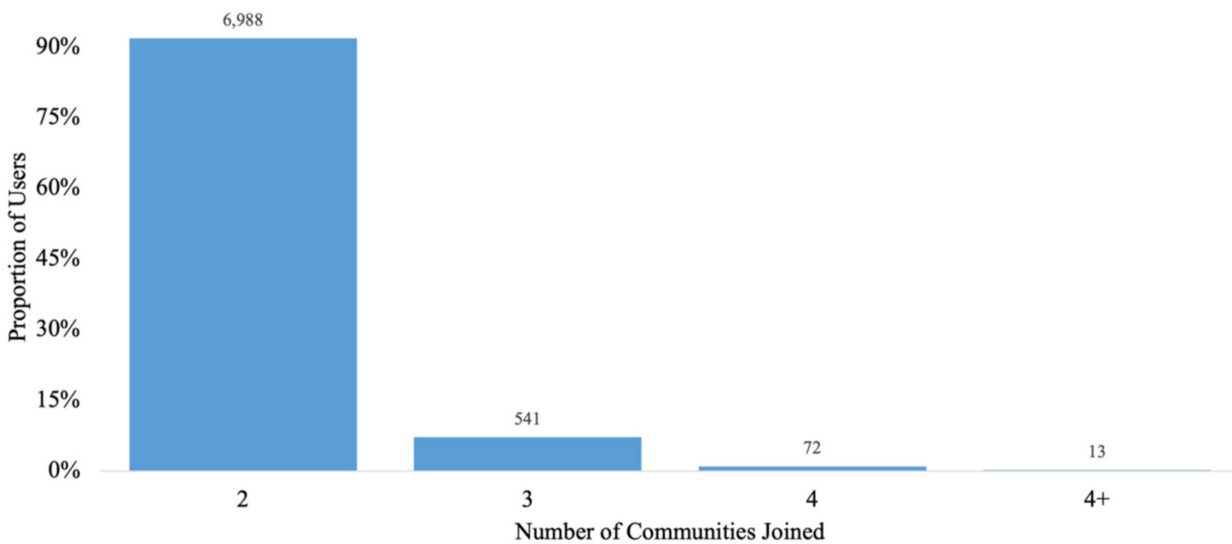


Figure O.1. Distribution of Multi-homed Users

Given the low incidence of multi-homing, we infer that its impact on cross-community variation in our empirical findings is likely minimal. Consequently, we believe that controlling for both time-level and community-level fixed effects would be adequate to mitigate concerns related to unobserved confounders in our analysis.

Appendix P: Network Exogeneity: Additional Evidence

In this section, we provide additional evidence to support the premise that the likelihood of underlying preference correlations among users, beyond their positions within the referral network, is minimal. While controlling for user-level fixed effects within our model would ideally be the most robust approach, such a control is impractical in our study. We are dealing with hundreds of thousands of users, and the majority have only made single-period decisions before permanently leaving the community, making the application of user fixed effects unfeasible.

To test the relative exogeneity of network connections, we proposed a hypothesis: If there are additional preference correlations between users, those who join through private channels would more likely exhibit such correlations between referrers and referees. To test this hypothesis, we compared the network effects experienced by users who joined through private channels to those who joined through public channels. This comparison is aimed at examining the potential presence of other underlying preference correlations.

Table P.1. Regression Results under Different Community Groups

Dependent Variable:	Renewal Decision (d)	
	Private Links	Website
	(1)	(2)
Price Effect		
Price	-0.035***	-0.117***
	(0.009)	(0.004)
Network Effect		
Referred	-0.061***	-0.063***
	(0.010)	(0.009)
Referrer's Decision	0.048***	0.045***
	(0.010)	(0.010)
Number of Referees	0.138***	0.144***
	(0.005)	(0.005)
User Characteristics	YES	YES
Community Characteristics	YES	YES
Observations	173,869	174,503
Joining Time FE	YES	YES
Renewal Time FE	YES	YES
Community FE	YES	YES
R^2	0.304	0.281

Notes: Standard errors are indicated in parentheses. Significance levels are denoted as follows: ∗p < 0.10, ∗∗p < 0.05, ∗∗∗p < 0.01.

Our dataset does not specify the individual entry method for each user into the platform. Instead, we have access to "join links" data at the community level, which reflects the proportion of referral links originating from private channels (e.g., direct messages on WeChat) versus public channels (e.g., web page advertisements). Using these extra data sources, we categorized communities into two groups: those with a high proportion of private channel referrals (above the median) and those with a higher ratio of public referral links (website-based). Despite potential overlaps between these groups, we conducted separate regressions for users in each group and reported the results in Table P.1. Our analysis revealed that the network effects estimated for the sample in Column (1)—users more likely to have unobserved preference correlations—and those for the sample in Column (2)—users less likely to have such correlations—are strikingly similar. These results indirectly but significantly mitigate our concerns about network endogeneity and suggest that our network connections are relatively exogenous.

Appendix Q: Empirical Evidence Supporting Myopic Assumption

This section presents additional empirical evidence supporting Assumption 1, which suggests that users make mutually exclusive renewal decisions in each period. Our dataset reveals that a user's decision to exit a community is essentially irreversible, with 99.65% of users who leave a community not returning. This quasi-irreversible nature implies that a straightforward regression between a user's decisions in consecutive periods would likely show a significant statistical relationship. Specifically, if $d_t = 1$, it implies $d_{t-1} = 1$.

To explore deeper, we adopted the following model specification: We calculated the community-specific renewal rate for each period and regressed it against the renewal rate from the same period in the previous year. The regression model is expressed as:

$$\bar{d}_{j,t} = \varrho_0 + \varrho_1 \bar{d}_{j,t-1} + \psi_j + \phi_t + \varepsilon_{j,t},$$

where $\bar{d}_{j,t}$ represents the renewal rate for community j at time t; ψ_j and ϕ_t are community-level and time-level fixed effects, respectively, with the latter at the year-month level.

Table Q.1. Community-Level Renewal Rate Analysis

Dependent Variable:	Community Renewal Rate ($\bar{d}_{j,t}$)			
	(1)	(2)	(3)	(4)
Lag Renewal Rate ($\bar{d}_{j,t-1}$)	0.308***	0.318***	-0.027	-0.025
	(0.022)	(0.021)	(0.023)	(0.022)
Observations	4,044	4,044	4,044	4,044
Time FE	NO	YES	NO	YES
Community FE	NO	NO	YES	YES
R^2	0.090	0.187	0.423	0.509

Notes: Standard errors are indicated in parentheses. Significance levels are denoted as follows: ∗p < 0.10, ∗∗p < 0.05, ∗∗∗p < 0.01.

Table Q.1 presents regression results across four models, examining the relationship between the current period's renewal rate ($\bar{d}_{j,t}$) and the previous year's same period renewal rate ($\bar{d}_{j,t-1}$). Each model controls for different fixed effects, as illustrated in Columns (2) to (4). The critical observation is the shift from Column (2) to Columns (3) and (4). Once community-level fixed effects are incorporated, the coefficient becomes insignificant. This finding indicates that the previous period's renewal probability within a community does not significantly impact the current period's renewal probability. The results provide empirical backing for our myopic assumption, reinforcing the notion that renewal decisions in each period are essentially made independently of past decisions.

Appendix R: Computation of Model Equilibrium

This section describes our methodology for computing the market equilibrium. Our model employs a tree-like network structure where the decisions of upstream users influence downstream users.

For a given community m, within its network $\mathcal{R}_{m,t}$ at time t, our initial step involves categorizing users in $I_{m,t}$ based on their referral hierarchy. The base level, $\mathcal{L}_0^m$, is defined as $\mathcal{L}_0^m = \{i \in I_{m,t} \mid \sum_{j \in I_{m,t}} r_{i,j} = 0\}$. This level consists of users who have not referred the community to others. The subsequent level, $\mathcal{L}_1^m$, is composed of users who have referred the community to individuals in $\mathcal{L}_0^m$, expressed as $\mathcal{L}_1^m = \{i \in I_{m,t} \mid \Pi_{j \in \mathcal{L}_0^m}(1 - r_{i,j}) = 0\}$. We denote the uppermost level as $\bar{m}$, where $\mathcal{L}_{\bar{m}}^m$ includes users who have referred the community to those in $\mathcal{L}_{\bar{m}-1}^m$, and have no referrers themselves, formally defined as $\mathcal{L}_{\bar{m}}^m = \{i \in I_{m,t} \mid \Pi_{j \in \bar{m}-1}(1 - r_{i,j}) = 0 \text{ and } \sum_{j \in I_{m,t}} r_{j,i} = 0\}$.

To compute the market equilibrium for community m, as delineated in Definition 1, we employ "backward induction and forward optimization," detailed as follows:

1. Starting with $\mathcal{L}_0^m$, we compute the optimal response for each i in this set. Their likelihood of opting for $d_{i,m,t} = 1$ is contingent upon upstream decisions, denoted by probabilities $p_{i,m,t,d}^1$ and $p_{i,m,t,d}^1$:

$$p_{i,m,t,d}^1 = \mathbf{P}\big(d_{i,m,t} = 1 | \{d_{j,m,t}\}\_(j \in \mathcal{L}_1^m)\big), \;\; p_{i,m,t,d}^0 = 1 - p_{i,m,t,d}^1.$$

   Notably, each i is associated with two probability sets: one for renewal/departure contingent on their referrer's renewal ($p_{i,m,t,1}^1 / p_{i,m,t,1}^0$) and another for renewal/departure based on their referrer's leave ($p_{i,m,t,0}^1 / p_{i,m,t,0}^0$).

2. With a predicted probability set $\{p_{i,m,t,1}^1\}_{i \in \mathcal{L}_0^m}$, we advance to $\mathcal{L}_1^m$. For all the $i \in \mathcal{L}_1^m$, we calculate the expected value $\mathbf{E}(\sum_{j \in \mathcal{L}_0^m} r_{i,j} d_{j,m,t} \mid d_{i,m,t} = 1)$, represented by $E_{i,m,t}^d$, as a function of $\{p_{i,m,t,1}^1\}_{i \in \mathcal{L}_0^m}$ and $d_{i,m,t}$. Given that $\mathbf{E}\big(d_{j,m,t} \big| d_{i,m,t}\big) = \mathbf{P}(d_{j,m,t} = 1 | d_{i,m,t})$, it follows that $E_{i,m,t}^d = \sum_{j \in \mathcal{L}_0^m} r_{i,j} p_{j,m,t,1}^1$.
3. Iteratively apply Steps 1 and 2 for each group $\mathcal{L}_k^m, k = 1, \dots, \bar{m} - 1$, calculating the choice probabilities for each user $i \in \mathcal{L}_k^m$ and compiling them as $\{p_{i,m,t,1}^1\}_{i \in \mathcal{L}_k^m}$. With this set of probabilities, we advance to the next tier, $\mathcal{L}_{k+1}^m$, to compute the values of $E_{i,t}^d$ for $i \in \mathcal{L}_{k+1}^m$.
4. For users in the top level $\mathcal{L}_{\bar{m}}^m$, we use expected values $\{E_{j,t}^d\}\_(j \in \mathcal{L}_{\bar{m}-1}^m)$ determine their utilities and decisions $\{d_{i,m,t}\}\_(i \in \mathcal{L}_{\bar{m}}^m)$.
5. The final step involves a forward optimization loop. For every tier $\mathcal{L}_k^m, k = \bar{m} - 1, \dots, 0$, we use the deisions $\{d_{i,m,t}\}\_(i \in \mathcal{L}_{k+1}^m)$ and $\{E_{i,t}^d\}_{i \in \mathcal{L}_k^m, d \in \{0,1\}}$ derived from the backward steps to finalize the decisions $\{d_{i,m,t}\}\_(i \in \mathcal{L}_k^m)$ for each level.

6. The complete market equilibrium for time t is then given by the union of decisions across all levels: $\mathcal{E}_{m,t} = \cup_{k \in \{0,\dots,\bar{m}\}} \{d_{i,m,t}\}$_$(i \in \mathcal{L}_k^m)$.

Appendix S: Validation of Estimation Method Using Simulation

This section details how we use simulation techniques to validate that our proposed estimation algorithm accurately recovers parameters from a specified Data Generation Process (DGP), drawing on methodologies akin to Su (2014). Our validation comprises three primary steps: setting initial parameters, generating sample sizes, and applying our estimation model to assess parameter recovery. The steps are as follows:

Data Generation Process (DGP) The procedure for generating simulated samples is outlined below:

1. Define the total number of users n with 30% being referred (i.e., $0.3 \times n$).
2. Create an empty $n \times n$ adjacency matrix.
3. Randomly select referred users from the total user pool.
4. Assign a non-referred user as a referrer to each referred user, updating the adjacency matrix while avoiding network cycles.
5. Generate observations based on the utility model, mirroring our model in the manuscript:

$$u(d_i = 1) = \beta_0 + \beta_p \ln \text{price}_i + \beta_R R_i + \beta_{RD} RD_i + \beta_r \mathbf{E}\left(\sum_{k \in I_{m,t}} r_{i,k} d_{k,t} / \sum_{k \in I_{m,t}} r_{i,k} | d_{i,t}\right) + \varepsilon_i^1;$$

where R_i indicates whether the user is referred by a referrer, RD_i represents the referees' decision and the expectation term the proportion of referees renewing their membership.

Simulation and Estimation The simulation process includes:

1. Set Initial Parameters: We define a set of utility function parameters as our starting point.
2. Generate Samples: Using the DGP, we create four sample sizes (100, 500, 1,000, and 3,000 users) to test our estimation method. Each sample assumes that 30% of users are referred, with renewal decisions based on predefined parameters.
3. Execute Estimation and Validation: Our estimation method is applied to each sample, and results are compared to the true parameters.

Table S.1 shows that our method consistently aligns well with the true parameters across different sample sizes. For large sample sizes, the precision of our estimates is satisfactory.

Robustness Check An additional simulation was conducted, modifying the initial parameter settings to assume a negative influence of downstream users on upstream users ($\beta_r < 0$) and $\beta_{RD} = 0$. This test aimed to determine if our method would erroneously identify positive correlations due to the DGP. Results in Table S.2 confirm the method's accuracy in retrieving the initial parameters under these conditions, further validating the reliability of our estimation in identifying network effect propagation paths.

Table S.1. Parameter Estimates for Simulated Samples ($\beta_r > 0$)

Dependent Variable:	Renewal Decision (d)				
	Initial Setting	$n = 100$	$n = 500$	$n = 1{,}000$	$n = 3{,}000$
	(1)	(2)	(3)	(4)	(5)
Price (β_p)	-2.00	-2.15	-1.91	-2.03	-2.09
		[-3.27, -1.09]	[-2.72, -1.12]	[-2.63, -1.45]	[-2.43, -1.75]
Referred (β_R)	-1.50	-1.61	-1.52	-1.71	-1.60
		[-4.65, 0.52]	[-2.84, -0.41]	[-2.37, -1.12]	[-1.97, -1.25]
Referrer's Decision (β_{RD})	2.00	3.20	1.94	2.25	2.24
		[0.79, 6.43]	[0.74, 3.34]	[1.60, 2.95]	[1.85, 2.64]
Expectation of Referees' Decisions (β_r)	3.00	3.36	3.70	3.46	3.27
		[1.45, 6.83]	[2.38, 5.43]	[2.69, 4.36]	[2.82, 3.76]
Constant (β_0)	12.00	12.97	11.41	12.06	12.35
		[6.71, 19.54]	[6.76, 16.21]	[8.62, 15.59]	[10.37, 14.37]

Notes: 95% confidence intervals are reported below estimates. 30% of the users in each group serve as referees.

Table S.2. Parameter Estimates for Simulated Samples ($\beta_r < 0$)

Dependent Variable:	Renewal Decision (d)				
	Initial Setting	$n = 100$	$n = 500$	$n = 1{,}000$	$n = 3{,}000$
	(1)	(2)	(3)	(4)	(5)
Price (β_p)	-2.00	-2.20	-1.88	-1.83	-2.11
		[-3.28, -1.16]	[-2.69, -1.09]	[-2.38, -1.31]	[-2.43, -2.11]
Referred (β_R)	-1.50	-1.57	-1.49	-1.59	-1.37
		[-2.66, 0.59]	[-2.17, -0.86]	[-1.98, -1.21]	[-1.58, -1.15]
Referrer's Decision (β_{RD})	0.00	-0.70	-1.58	0.26	-0.26
		[-3.74, 1.33]	[-3.49, -0.19]	[-0.30, 0.80]	[-0.61, 0.09]
Expectation of Referees' Decisions (β_r)	-3.00	-7.73	-10.16	-1.93	-3.76
		[-13.58, -3.40]	[-15.26, -6.24]	[-3.26, -0.64]	[-4.54, -3.01]
Constant (β_0)	12.00	13.31	11.42	10.99	12.67
		[7.27, 19.71]	[6.78, 16.24]	[7.87, 14.17]	[10.81, 14.56]

Notes: 95% confidence intervals are reported below estimates. 30% of the users in each group serve as referees.

Appendix T: Average Marginal Effects Analysis

While the primary results from our structural model are detailed in Table 12, this section aims to shed light on the average marginal effects of changes in price and network-related variables on user decisions. We have employed both logit and probit models for this analysis, and the outcomes are collated in Table T.1.

Table T.1. Logistic/Probit Regression Results (Average Marginal Effects)

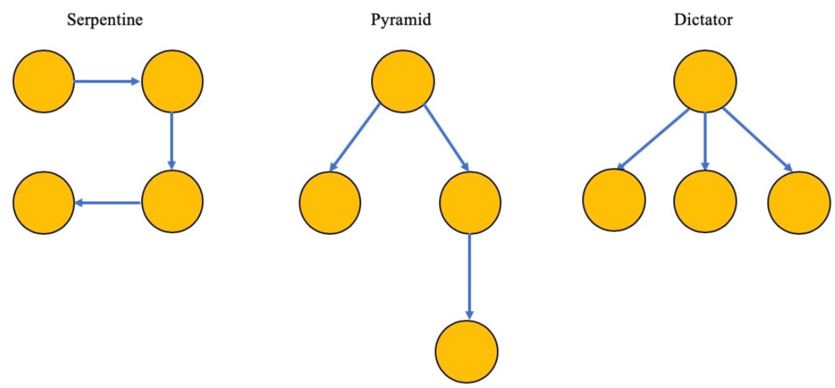
Dependent Variable:	Renewal Decision (d)			
	Logit	Probit	IV Logit	IV Probit
	(1)	(2)	(3)	(4)
Price Effect				
Price	-0.092***	-0.092***	-0.234***	-0.226***
	(0.004)	(0.004)	(0.006)	(0.006)
Network Effect				
Referred	-0.042***	-0.040***	-0.041***	-0.039***
	(0.007)	(0.007)	(0.007)	(0.007)
Referrer's Decision	0.022***	0.020***	0.022***	0.020***
	(0.007)	(0.007)	(0.007)	(0.007)
Number of Referees	0.245***	0.224***	0.243***	0.223***
	(0.008)	(0.007)	(0.008)	(0.007)
User Characteristics	YES	YES	YES	YES
Community Characteristics	YES	YES	YES	YES
Control Function				
$\hat{\eta}$			0.001***	0.001***
			(0.000)	(0.000)
Observations	263,161	263,161	263,161	263,161
Joining Time FE	YES	YES	YES	YES
Renewal Time FE	YES	YES	YES	YES
Community FE	YES	YES	YES	YES
Log Likelihood	-129,988.950	-129,986.500	-129,472.930	-129,491.090

Notes: Standard errors are indicated in parentheses. Significance levels are denoted as follows: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

Our computation of the average marginal effects reveals that the results from the logit and probit models are strikingly similar. This consistency indicates that, despite potential differences in the coefficient values directly derived from each model, their impact at an average level remains uniform across both models. Essentially, this finding underscores the robustness of our study's findings, showing that the choice of either logit or probit model does not significantly alter the interpretation of the average effects of price and network variables on user decisions.

Appendix U: Additional Evidence and Implications for Marketing Strategy

Our structural estimations indicate that discount policies are particularly effective in networks characterized by a high beta index, indicative of enhanced network connectivity, which in turn significantly enhances user retention. Prior studies on network effects has primarily focused on the influence of peers or the size of the network as indicated by the beta index (e.g., Ackerberg and Gowrisankaran 2006; Ryan and Tucker 2012). However, it is crucial to note that the beta index might not capture the full extent of network effects on outcome variables, especially when network decisions are made endogenously. To illustrate this point, Figure U.1 presents three different network models: Serpentine, Pyramid, and Dictator. Although these models share the same beta index values, their levels of centralization, as measured by closeness, progressively increase from the Serpentine to the Dictator model. Under the pricing system of our platform, these three structures would generate the same revenue in the initial period. However, their impact on renewal rates in subsequent periods could vary significantly. This observation implies that if platform owners are primarily concerned with single-period revenue, they might not show a preference for any particular network structure. In contrast, if their focus extends to multi-period revenue generation, certain network structures might become more preferable.



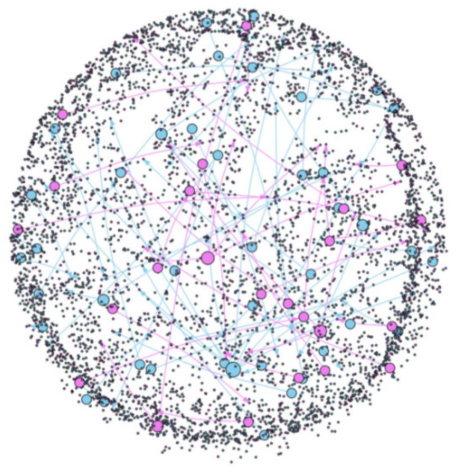
Notes: The three network structures in the above figure all have the same level of beta index and number of nodes, but the centralization increases once from left to right. We refer to their structures as Serpentine, Pyramid, and Dictator, respectively.

Figure U.1. Different Network Structures with the Same Level of Connectivity

In this section, we explore which network structures might be more desirable for platforms aiming to foster long-term user engagement. Additionally, we discuss how platforms can encourage user-to-user referrals to establish networks that foster higher renewal rates.

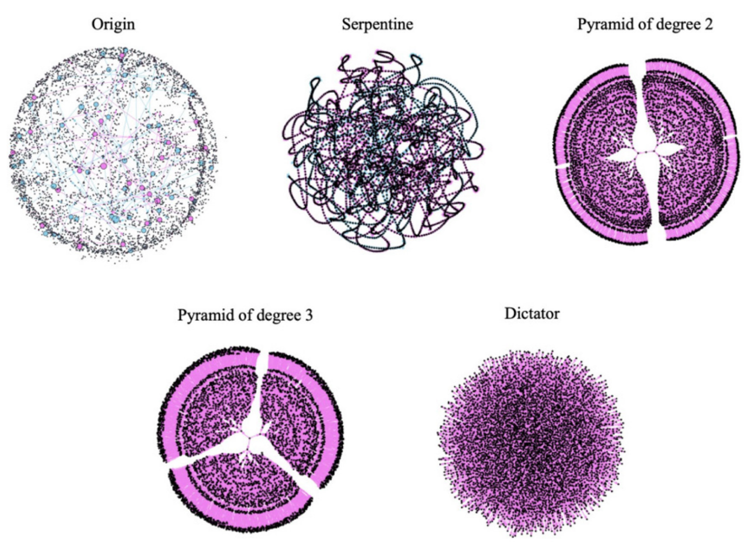
U.1 Connectivity and Centrality Analysis

To examine the impact of connectivity and centrality on renewal rates, we conducted counterfactual studies using various network structures. These studies focus on two key variables—connectivity and centrality—and their influence on average renewal rates. We use the community depicted in Figure U.2 as a base for reconfiguring its network under different structures: Serpentine, Pyramid (with degrees 2 and 3), and Dictator. In these models, each new member is assumed to recommend the next user(s) in a sequence, with the Dictator model having a single user referring all others. Figure U.3 illustrates these networks, maintaining a constant beta index but varying degrees of centrality.



Notes: Each vertex in the above network diagram represents a user, with pink color for users who have joined in the last year and blue for users who have made at least one renewal decision. The vertex size represents the rewards that users receive through referrals, and the arrows represent the referral relationships between users. Among the 5,792 users, 4,057 are users who have made at least one renewal decision (“old users”) and 1,735 are users who joined the community in the past year (“new users”).

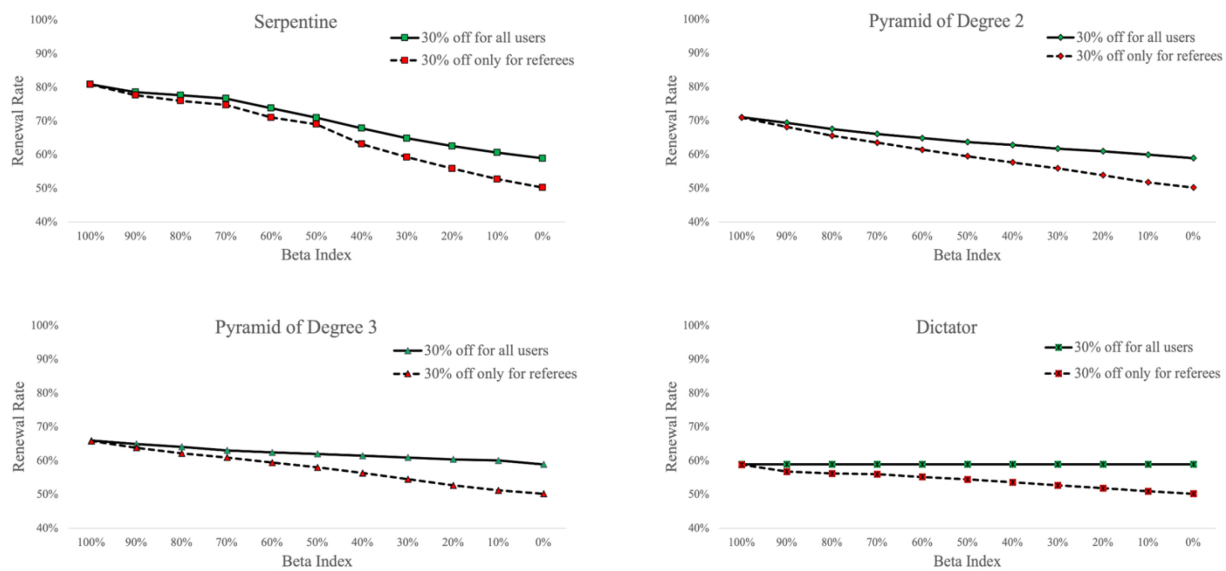
Figure U.2. An Illustration of a Referral Network in the Data



Notes: In the figure above, we show five different network structures based on the community in Figure T.2: the original network structure on the upper left, the Serpentine network structure on the upper middle, the Pyramid network structure of degree 2 on the upper right, the Pyramid network structure of degree 3 on the bottom left, and the Dictator network structure on the bottom right.

Figure U.3. Network Structures with the Same Connectivity and Different Centrality Degrees

Figure U.4 illustrates the effect of the beta index on renewal rates under a 30% discount policy. This analysis compares the impacts of a uniform 30% discount applied to all users versus a discount offered exclusively to referees across various network structures. In a fully connected network, represented by a beta index of 100%, every user is linked to every other user. The renewal rates depicted in the graph for different network structures under a uniform 30% discount are as follows: The Serpentine network shows a rate of 80.90%, Pyramid degree 2 has 71.01%, Pyramid degree 3 at 65.95%, and the Dictator network at 58.91%. These rates decrease as connectivity, indicated by the beta index, decreases. Specifically, for every 10% reduction in the beta index, there is a corresponding decrease in renewal rates by 3.20% for the Serpentine, 1.21% for Pyramid degree 2, and .70% for Pyramid degree 3 networks. In the Dictator model, where there is only a single referrer with multiple referees, the reduction in connectivity has a negligible impact on renewal rates.



Notes: Our counterfactual analysis explores variations in renewal rates across distinct network structures, accommodating a consistent user base of 5,792 individuals. We analyze four specific configurations: the Serpentine network (upper left), the Pyramid network with a degree of 2 (upper right), the Pyramid network with a degree of 3 (bottom left), and the Dictator network (bottom right). The beta index and renewal rates are plotted along the horizontal and vertical axes, respectively, for these network structures. The beta index is assessed across a spectrum from 0% to 100%, increasing in increments of 10%. At each specified beta index level—take for instance, a beta index of .9—we remove the last 10% of connections from the original network and recompute the market equilibrium for the truncated network structure. A beta index of 0 signifies a scenario devoid of any network effects. The impact on renewal rates is depicted in two distinct manners: A solid line with green dots indicates the renewal rate effect when all users are granted a 30% discount on the renewal price, illustrating the response to a uniform pricing strategy. Conversely, a dashed line with red dots delineates the effect when only referees benefit from a 30% discount on their renewal price, demonstrating the outcomes of a targeted discount policy. Each line tracks the respective changes in renewal rates in relation to shifts in the beta index, thereby quantifying the influence of network structure on the efficacy of pricing strategies.

Figure U.4. Beta Index and Corresponding Renewal Rates

Moreover, the trend across these network structures shows a nearly linear decrease in renewal rates as the centrality of the network decreases. When the beta index is reduced to zero, indicating minimal connectivity, the renewal rates across all network types converge to approximately 55.42%. Under a policy that targets referees with discounts, the trend of decreasing renewal rates is similar to that of the uniform discount policy, but the overall renewal rates are lower due to the narrower scope of the discount policy. These findings suggest that for a referral network to efficiently maximize the impact of price discount policies on renewal rates, it should aim for high connectivity while keeping the degree of centrality as low as possible.

To further show the relative advantage in renewal rates associated with a lower degree of closeness- based centrality, we held the level of connectivity constant, as measured by the beta index, and varied the degree of network centralization to observe its impact on renewal rates. The findings from these analyses are summarized in Table U.1 and reveal important insights about the role of network centralization. Our results demonstrate that as the degree of network centralization increases, the overall renewal rate tends to decrease, despite a consistent level of connectivity across different network structures. Specifically, the Serpentine network, characterized by a one-to-one referral structure, showed the highest retention rate. In contrast, the Dictator network, which represents a highly centralized structure where one user refers all others, yielded a significantly lower user retention rate of 47.91%, only slightly better than the baseline scenario. When comparing different network structures, the study observed a diminishing marginal decrease in renewal rates with increasing centralization. For instance, the renewal rate drops by 9.30% when moving from a Serpentine network to a Pyramid network of degree 2. However, further increases in centralization (from Pyramid degree 2 to Pyramid degree 3, and then to the Dictator model) result in smaller reductions in renewal rates.

Table U.1. Network Centralities and Corresponding Renewal Rates

Network Structure	Serpentine	Pyramid (degree 2)	Pyramid (degree 3)	Dictator
Renewal Rate (%)	70.80	61.50	57.30	47.91
Number of Nodes	5,792	5,792	5,792	5,792
Number of Edges	5,791	5,791	5,791	5,791
Beta Index	.999	.999	.999	.999
Closeness Centrality	.00030	.084	.128	1

Notes: The Beta Index takes the ratio of the number of edges over the number of nodes. The Centrality degree (CC) of the graph in terms of closeness is designed by Leavitt (1951): $C_C = \frac{\sum_{i=1}^{n}[C_C'(p^*) - C_C'(p_i)]}{[(n^2-3n+2)/(2n-3)]}$, where n denotes the number of points, $C_C'(p_i)$ denotes the point centrality of a point p_i (Beauchamp (1965) suggested using $C_C'(p_i) = \left[\frac{\sum_{i=1}^{n} d(p_j,p_i)}{n-1}\right]^{-1}$, where $d(p_j,p_i)$ denotes the number of edges in the geodesic connecting p_j and p_i), and $C_C'(p^*)$ denotes the largest value of $C_C'(p_i)$.

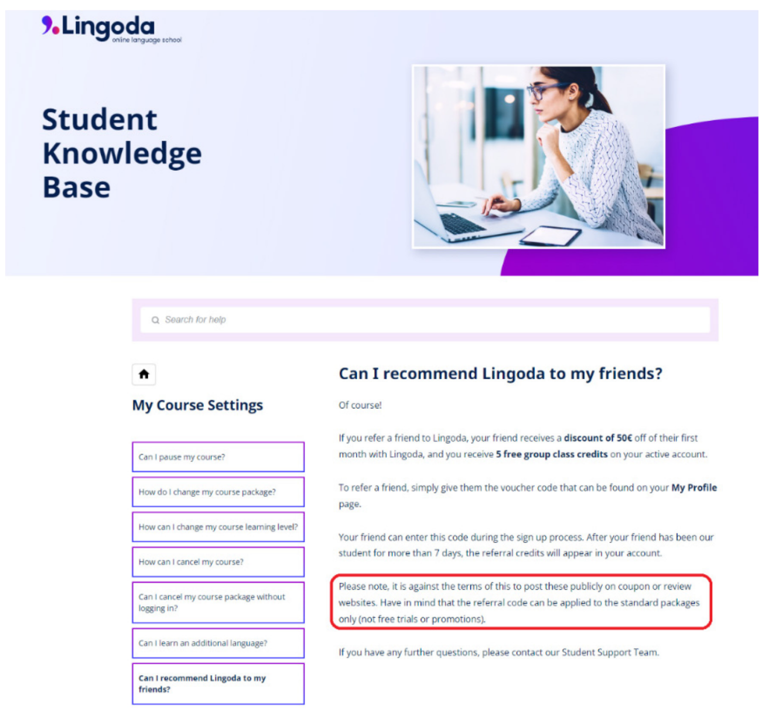
These insights challenge the traditional focus in business strategies, where the emphasis has often been on leveraging super influencers for user acquisition. Our findings suggest an alternative approach, where incentivizing each user to sequentially refer others in a chain-like manner could be more effective in enhancing long-term user retention. A less centralized network structure seems to maximize the positive impacts of referral programs by allowing a wider spread of influence through the network. Our study reveals the importance of a balanced approach in crafting referral policies, where platform owners must weigh the trade-offs between network connectivity and centrality. The subsequent sections of our analysis further explore this aspect and provide additional empirical evidence to support our conclusions.

U.2 Anecdotal Evidence of Referral Policy Design

Our empirical analysis highlights the significant impact of referral programs on user renewal decisions. Our counterfactual studies suggest that in a highly connected network, a lower degree of centrality could enhance renewal rates. This section aims to offer insights for business managers on balancing connectivity and centrality in referral policies. The counterfactual analysis in Section reveals that while increasing connectivity in referral networks is crucial, decentralization, equating to the quality of network links, is equally important. In less centralized networks, users typically form one-to-one referral relationships, suggesting personal recommendations within close social circles. How platforms structure referral networks depends largely on their referral strategies.

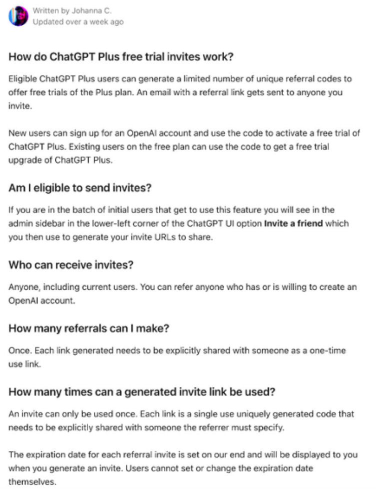
Platforms adopt diverse referral policies, often balancing centrality and connectivity based on the nature of their services. In this section, we illustrate two distinct types of referral policies based on real-world anecdotal evidence.

Referral Policies Balancing Centrality and Connectivity Platforms like Lingoda encourage personal referrals within close networks, offering incentives while discouraging public sharing of links to limit network centralization. This approach results in more personal, one-to-one referrals and reduces overall network centrality. Figure U.5 illustrate the referral policy of the platform. OpenAI's ChatGPT also follows a similar strategy by restricting users to a maximum of three referral links, each usable by only one other user, to prevent the dominance of super-influencers and maintain a lower level of network centrality. The referral policy of ChatGPT is illustrated in Figure U.6.



Notes: Source: https://lingoda-students.elevio.help/en/articles/250-can-i-recommend-lingoda-to-my-friends, accessed Nov 2022.

Figure U.5. Referral Policy from Lingoda



Written by Johanna C.
Updated over a week ago

How do ChatGPT Plus free trial invites work?

Eligible ChatGPT Plus users can generate a limited number of unique referral codes to offer free trials of the Plus plan. An email with a referral link gets sent to anyone you invite.

New users can sign up for an OpenAI account and use the code to activate a free trial of ChatGPT Plus. Existing users on the free plan can use the code to get a free trial upgrade of ChatGPT Plus.

Am I eligible to send invites?

If you are in the batch of initial users that get to use this feature you will see in the admin sidebar in the lower-left corner of the ChatGPT UI option **Invite a friend** which you then use to generate your invite URLs to share.

Who can receive invites?

Anyone, including current users. You can refer anyone who has or is willing to create an OpenAI account.

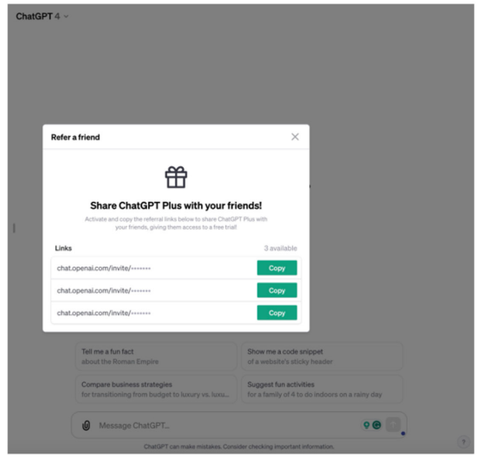
How many referrals can I make?

Once. Each link generated needs to be explicitly shared with someone as a one-time use link.

How many times can a generated invite link be used?

An invite can only be used once. Each link is a single use uniquely generated code that needs to be explicitly shared with someone the referrer must specify.

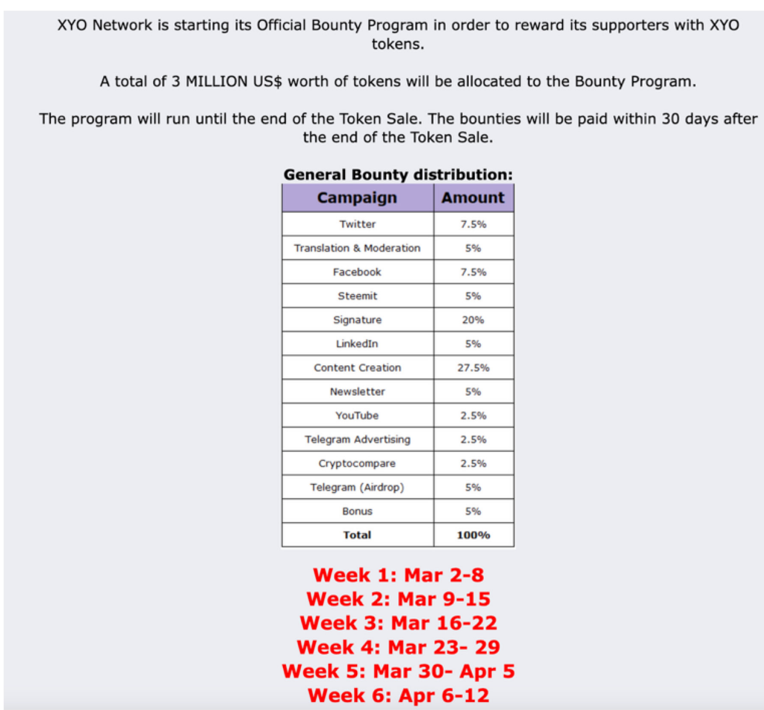
The expiration date for each referral invite is set on our end and will be displayed to you when you generate an invite. Users cannot set or change the expiration date themselves.



Notes: Source: https://help.openai.com/en/articles/8381046-free-trial-invites-faq, accessed Nov 2023.

Figure U.6. Referral Policy from ChatGPT

Referral Policies Prioritizing Connectivity Contrasting the first type, some platforms, especially in the ICO sector, prioritize building highly connected networks. The XYO Network, for instance, ran a high-reward referral campaign during its ICO phase to quickly attract super-influencers. Such campaigns offer significant rewards with few restrictions, focusing on rapid and widespread user acquisition, leading to highly centralized networks. The example is illustrated in Figure U.7.



XYO Network is starting its Official Bounty Program in order to reward its supporters with XYO tokens.

A total of 3 MILLION US$ worth of tokens will be allocated to the Bounty Program.

The program will run until the end of the Token Sale. The bounties will be paid within 30 days after the end of the Token Sale.

General Bounty distribution:

Campaign	Amount
Twitter	7.5%
Translation & Moderation	5%
Facebook	7.5%
Steemit	5%
Signature	20%
LinkedIn	5%
Content Creation	27.5%
Newsletter	5%
YouTube	2.5%
Telegram Advertising	2.5%
Cryptocompare	2.5%
Telegram (Airdrop)	5%
Bonus	5%
Total	**100%**

Week 1: Mar 2-8
Week 2: Mar 9-15
Week 3: Mar 16-22
Week 4: Mar 23- 29
Week 5: Mar 30- Apr 5
Week 6: Apr 6-12

Notes: Source: https://bitcointalk.org/index.php?topic=3053801.0, accessed Nov 2023.

Figure U.7. Referral Policy from XYO Network

These above examples illustrate how platform-specific goals and contexts influence the design of referral policies and the resultant network structures. Platforms must carefully consider their long-term objectives and user dynamics when crafting these policies.

U.3 Further Analytical Evidence

Beyond anecdotal evidence, we seek to substantiate these perspectives with empirical study in this section. In our empirical analysis, we explore the tendency of community owners to balance connectivity and centrality within their networks. Our dataset encompasses several hundred different communities, each with varying levels of connectivity and centrality, providing us the opportunity to test the following hypotheses:

Hypothesis 1. *A larger user base correlates with a lower level of network centrality*.

Hypothesis 2. *More referral relationships within personal networks correlate with a lower level of network centrality*.

To explore the nature of referral relationships within our dataset, we obtained additional community-level data from the platform that allowed us to categorize referral links based on their dissemination channels. This categorization helps us understand the varying degrees of personal connection between referrers and referees. The primary channels identified include chat software (such as WeChat and QQ), social platforms (like Weibo), and hyperlinks shared across websites. Our analysis distinguishes the referral links as follows:

1. Chat Software (WeChat and QQ): We hypothesize that links shared through these platforms are indicative of personal connections, as they are commonly used for communication with close friends and relatives. This mode of sharing suggests intimate, one-to-one social ties and a tendency towards less centralized network structures. Later, we use “Private Links” to denote these links sent via chat software.
2. Social Platforms (e.g., Weibo): Links shared on these platforms likely represent less intimate relationships, as social platforms are generally used for broader, less personal interactions. This could imply weaker information matching and social ties, contributing to more centralized network structures.
3. Hyperlinks on Websites: This method is the most public and impersonal, suggesting the least intimate form of referral and potentially highly centralized networks.

Our findings show that a significant portion of referral links, accounting for 59.2% (572,345 out of 966,717), were disseminated via personal chat software. This high proportion indicates a predominant reliance on personal social networks, such as family and friends, for referrals. Table U.2 presents a detailed proportional distribution of each type of referral link, shedding light on the various referral strategies employed across different communities.

In our study, we use the new data sources to calculate network metrics for each community, specifically focusing on the proportion of private referral links (shared via chat software channels). Our analysis aimed to investigate the relationship between the community size, the nature of referral links, and the network structure’s centrality and connectivity. The findings, presented in Table U.3, are aligned with the above hypotheses. Should Hypothesis 1 holds true, as the number of users in a community increases, there should be a decrease in centrality alongside an increase in connectivity. This implies that owners of larger communities might prioritize the quality of referrals, thus opting for a less centralized network structure. Should Hypothesis 2 holds true, an increase in the ratio of private links (indicative of more personal, targeted recommendation) should correlate with a decrease in network centrality.

Table U.2. Different Type of Links Sent from Communities

Statistic	M	SD	Min	Max	Proportion (%)
Chat Software	727.20	3,640.71	.00	54,802.00	59.19
Social Platform	12.39	82.53	.00	1,622.00	1.02
Hyperlink	488.77	2,160.20	.00	32,454.00	39.78

Notes: A total of 966,717 links were shared from the community between 2016 and 2021, with 59.2% transmitted via chat software. These summary statistics were derived from annual community data. “Chat Software” encompasses WeChat and QQ, while “Social Platform” refers to Weibo. “Hyperlink” is categorized into “Invite-URL” (direct sharing of invitation links) and “Copy-URL” (copying invitation links to share with others).

Table U.3. Community-level Centrality and Connectivity

Dependent Variable:	*Closeness-based Centrality*					
	(1)	(2)	(3)	(4)	(5)	(6)
Connectivity	3.991e-06***	3.862e-06***	3.858e-06***	5.775e-06***	5.610e-06***	5.594e-06***
	(1.040e-06)	(1.059e-06)	(1.065e-06)	(1.484e-06)	(1.534e-06)	(1.537e-06)
Number of Users	−7.102e-09	2.073e-08	2.087e-08			
	(3.758e-07)	(4.355e-07)	(4.361e-07)			
Connectivity × Number of Users	−7.821e-07**	−7.900e-07**	−7.888e-07**			
	(3.355e-07)	(3.384e-07)	(3.399e-07)			
Private Links				−6.059e-07	5.651e-07	4.792e-07
				(2.788e-06)	(3.790e-06)	(3.810e-06)
Connectivity × Private Links				−6.600e-06***	−6.380e-06***	−6.369e-06***
				(2.374e-06)	(2.443e-06)	(2.446e-06)
Observations	350	350	350	350	350	350
Year FE	NO	YES	YES	NO	YES	YES
Community Type	NO	NO	YES	NO	NO	YES
R^2	.041	.047	.047	.051	.051	.051

Notes: ∗p < .1; ∗∗p < .05; ∗∗∗p < .01. Standard errors are reported in parenthesis. The ratio of private links is calculated using the number of links sent through chat software divided by the total number of links sent in each year. “Community Type” is assigned a value of 1 for science-oriented communities and 0 for those focused on economics. Communities with zero referral links were excluded, resulting in a sample of 350 valid observations.

The results align closely with our expectations. Initially, we discover a positive correlation between high centrality and high connectivity: The more users join via referral links, the more likely it is for a community to host super influencers, culminating in a more centralized network. Subsequently, our findings reveal that network centrality is negatively associated with both the total number of users and the proportion of private referral links. The data, as shown in Columns (1) to (3), consistently suggest a plausible explanation that as communities grow in size, their owners tend to promote less centralized networks. This trend suggests a strategic shift from enhancing connectivity and centrality through super influencers to promoting sustainable growth via a less centralized yet significantly connected referral architecture. Further corroborating this trend, Columns (4) to (6) reveal that with an increase in the proportion of privately shared referral links, network centralization tends to decrease. This pattern underlines the effective role of personal network link sharing in fostering referral networks with lower centrality.

References

Ackerberg, D. A. and Gowrisankaran, G. (2006). Quantifying equilibrium network externalities in the ACH banking industry. *The RAND Journal of Economics*, *37*(3):738-761.

Beauchamp, M. A. (1965). An improved index of centrality. *Behavioral Science*, *10*(2):161-163.

Leavitt, H. J. (1951). Some effects of certain communication patterns on group performance. *The Journal of Abnormal and Social Psychology*, *46*(1):38.

Ryan, S. P. and Tucker, C. (2012). Heterogeneity and the dynamics of technology adoption. *Quantitative Marketing and Economics*, *10*(1):63-109.

Su, C. L. (2014). Estimating discrete-choice games of incomplete information: Simple static examples. *Quantitative Marketing and Economics*, *12*(2):167-207.