Platform Leakage: Incentive Conflicts in Two-Sided Markets

Yingkang Xie*

Northwestern University†

Huaiyu Zhu

Lalamove (Huolala)

August 1, 2022

Abstract

Leakage happens when buyers and sellers coordinate outside the platform to cut out the middleman, usually to avoid paying fees. Although platforms are concerned about losing revenue, leakage—by its very nature—is hard to measure and manage. Using geolocation data from a large on-demand service platform for cargo delivery, we identify offline transactions that are typically hard to track in online marketplaces. We exploit a quasi-experiment that gradually introduced driver commissions, thereby generating variation in participants' incentives for leakage. The introduction of this commission increased leakage by nearly four percentage points, doubling the percentage of offline transactions we detected. We leverage the variation in commission fees to estimate price sensitivities and transaction costs in a structural model. The likelihood of leakage increases as the quoted price of the delivery increases, as the drivers' potential savings in the commission exceed the costs of offline coordination. Our model estimates suggest that customers typically receive half of the commission savings from drivers to rationalize their agreement to leakage. Counterfactuals show that a stronger bargaining power of customers would exacerbate platform leakage. To conclude, we discuss how targeting coupons, monitoring technology, and strategic matching policy can mitigate leakage by aligning incentives, which are alternatives to ex-post punishments.

Keywords: disintermediation, platforms, incentives, transaction costs, bargaining.

*Kellogg School of Management, Evanston, IL 60208. Email: yingkang.xie@kellogg.northwestern.edu.

†I want to thank my committee chair Brett Gordon for his exceptional mentorship and constant support. I also thank Eric Anderson, Hemant Bhargava, Kent Grayson, Gaston Illanes, Lakshman Krishnamurthi, Tai Lam, Angela Lee, Ilya Morozov, Daniel Spulber, Artem Timoshenko, Anna Tuchman, Caio Waisman, Jingyi Wang, Joonhyuk Yang, and the seminar participants at Kellogg for their insightful comments.

1 Introduction

Platform businesses help buyers and sellers find each other and engage in convenient and trustworthy transactions (Einav et al., 2016). However, platforms face the challenge of disintermediation – buyers and sellers can transact directly outside the platform to circumvent the platform fees. For example, Uber and Lyft drivers may ask clients to cancel the trip on the app and pay them offline to avoid the commission (Bellotti et al., 2017). Offline transactions are known as "leakage" (Hagiu and Wright, 2022; Ladd, 2022). A Harvard business case documents that approximately 90% of transactions started in a freelance marketplace are conducted offline (Zhu et al., 2018). Given the potential loss of revenue, platforms are highly motivated to monitor leakage and design incentives to minimize it. However, by its very nature, leakage is hard to measure, which likely explains the limited empirical analyses on this topic despite the substantial interests from practitioners and theorists[1].

This paper uses proprietary data from China's largest on-demand cargo delivery platform. The unique data allow us to identify disintermediated transactions, characterize the extent of leakage, and assess how leakage may vary in response to changes in platform fees. The data include a pricing experiment with a staggered rollout design that launched a driver-side commission with a 15% fee in different cities at different times. The average cancellation rates of the 33 treatment cities went up by about 5% (from 23.57% to 28.54%) after the drivers were charged a fee. The potential extent of leakage motivated the development of detection technology. Our novel detection algorithm (Xie et al., 2022) combines both geolocation and job cancellation data to flag[2] disintermediation at the transaction level. As far as we know, our work is one of the few, if not only, studies that uses a direct measure of disintermediation, rather than indirect measures such as the intentions to disintermediate or the reduction in platform engagement (Gu and Zhu, 2021; Zhou et al., 2022). We then estimate a structural

[1]Theorists study disintermediation with analyses on leakage in online marketplaces (Chaves, 2018; He et al., 2020; Hagiu and Wright, 2022; Peitz and Sobolev, 2022) and showrooming in retail (Wu et al., 2004; Balakrishnan et al., 2014; Jing, 2018; Kuksov and Liao, 2018; Mehra et al., 2018; Wang and Wright, 2020)

[2]A transaction was disintermediated if (1) the assigned driver passed the origin and destination of the canceled trip around the time of service, and (2) the customer did not request and complete similar trips.

model to quantify the underlying factors that motivate or discourage leakage, leveraging the quasi-experimental variation in driver-side fees and customer-side coupons. Our estimates inform new platform designs that can mitigate leakage.

We provide insights into proactive retention, product design, and matching algorithms in two-sided markets. These preventive measures are ex-ante alternatives to the ex-post punishments that are common in the industry. For example, many platforms threaten to ban accounts that initiate off-platform transactions (Uber, 2022; Airbnb, 2022a; eBay, 2022a). However, it is unclear whether platforms have efficient ways to recover all offline transactions and verify who in the buyer-seller pair is at fault for proposing leakage, given that coordination typically happens under the table. Moreover, platforms that rely on punishments may not only antagonise users but also lose future revenue from banned accounts[3]. Lastly, recurring expenses are likely to incur for policy enforcement in the game of fraud detection[4] where people learn to hide their activities from the platform to avoid punishments.

A deeper understanding of the incentive mechanisms affecting leakage can help platforms identify the appropriate economic levers to prevent disintermediation before it happens. Drivers and customers are less likely to collude if their offline transaction costs outweigh the part of commission savings they each retain or receive after bargaining. Our model estimates suggest that the platform may want to allocate marketing efforts toward drivers who are sensitive to commission. In contrast, giving coupons to customers might not be an effective tool to reduce leakage. To encourage customers to stay, the platform may develop services[5] to provide standalone value or reduce the costs of on-platform transaction. Lastly, to prevent leakage, the platform can strategically match drivers and customers such that their joint costs of offline transactions are larger than the fees they pay to the platform.

Since customers were not charged any platform fees, drivers might offer them a discount

[3]In our random sample, 2/3 of drivers were involved in at least one disintermediated transaction over 137 days. It is impossible to kick everyone out. See Section 3.3 Descriptive Statistics for more details.

[4]With any technology becoming weaponized, there is an inevitable race between the countermeasures to that technology and the development of counters to the countermeasures.(Williamson and Scrofani, n.d.)

[5]For example, the platform can grant customers access to the remote camera in cargo space to monitor their goods in transit. This reduces the need for customers to send someone to accompany the goods.

for offline transactions. To better understand how drivers and customers (two parties) might share the commission fees (surplus) recouped from the platform (intermediary), we adopt a commonly used solution concept from the bargaining literature (Nash, 1950; Sieg, 2000; Zhang et al., 2021; Jiang, 2022) to microfound our model. We find that the two parties split the commission savings in half, on average, to rationalize the joint decision of leakage in our sample. Drivers are typically the initiators of leakage and provide a take-it-or-leave-it offer[6] to customers to make them indifferent about where to transact. Our counterfactuals show that the likelihood of leakage is higher when customers have stronger bargaining power. Since the bargaining power depends on whether the parities have outside options (Backus et al., 2020) and sufficient time (Rubinstein, 1982) for negotiation, a focus on fast and last-minute matching might help the platform mitigate leakage.

Our model estimates also tell us which and when platform services[7] can justify the fees of having the platform as the guardian of trust (Shapiro, 1987). We find that the cost is about ¥6 ($1) for an average driver-customer pair to give up the digital escrow payment service provided by the platform. Moreover, suggestive evidence shows that customers are more likely to give up the convenience of tracking drivers' location in the app if they decide to send someone[8] to accompany the goods in transit. In total, the average offline transaction cost for a typical transaction is between ¥20($3) and ¥25($4), and is higher than the average commission fee of ¥16.5($2.5) the platform receives. These estimates not only help us to understand the value of the platform, but also provide the basis for the platform to evaluate alternative pricing strategies (e.g., a higher commission rate) and potential investment opportunities in new products (e.g., cargo monitoring technology).

Our investigation contributes to the literature of two-sided markets. We not only provide direct evidence that leakage exists, but also show that leakage is subject to heterogeneity in price sensitivity and transaction costs across the two sides of the market. Researchers in

[6]When drivers have full bargaining power, customers only get the minimum transfer that make them indifferent between on-platform and off-platform transaction. See Appendix G for more details.

[7]Platforms provide escrow, transaction monitoring, and dispute settlements (Edelman and Hu, 2016).

[8]Having one(two) passenger(s) is associated with lower offline frictions by roughly ¥1 (¥2).

industrial organization often assume these problems away when analyzing fees and subsidies (Rochet and Tirole, 2006; Weyl, 2010). Without taking into account leakage, platforms may fail to maximize profits at equilibrium, because optimal fees and subsidies depend not only on the price elasticity but also on transaction costs (Spulber, 2019; Hagiu and Wright, 2022). We conduct one of the few, if not only, empirical studies that take into account both differential price elasticities and heterogeneous transaction costs to guide platform design. We quantify transaction costs (Coase, 1937; Williamson, 1987) that prevent individuals from coordinating without the platform due to the hassle, inconvenience, and additional efforts.

To the best of our knowledge, this is the first empirical work to study the effect of platform fees on leakage with discussion about its boundary conditions. The most relevant studies to our work are by Gu and Zhu (2021) and Zhou et al. (2022), which investigate how disintermediation increases with trust achieved by reputation systems or repeated interactions. They focus on continuous transactions and do not study how monetary incentives play a role in leakage. Our application in on-demand services demonstrates that disintermediation can happen even in one-off transactions, as long as the commission savings exceed the costs of offline transactions. Our research in the gig economy[9] can motivate new strategies of other businesses (e.g., retail, e-commerce, the sharing economy) that involve buyers and sellers who make decisions on a daily basis about whether or not to engage in direct sales.

The rest of the paper proceeds as follows. Section 2 describes the background on the on-demand logistics platform, and demonstrates the preliminary evidence of leakage that motivates the development of detection algorithms. Section 3 details our direct measure of disintermediation, characterizes the extent of leakage, and summarizes the data that informs the identification for our structural model. In Section 4, we set up the model that accounts for leakage responses to changes in incentives. Section 5 presents the model estimates and Section 6 discusses their implications for platform design. Section 7 concludes by summarizing the findings and limitations with discussion on future research on this topic.

[9] The gig economy connects customers with independent contractors who work on their own schedules.

2 Empirical Setting

Disintermediation undermines the platform's ability to capture the value it creates (Ladd, 2022). Money leaks due to disintermediation are difficult to discover and curtail, so practitioners call this challenge "platform leakage." One key consideration for the platform's success is the extent of the leakage and how they may deal with it (Hagiu and Wright, 2021).

The extent of leakage varies widely across marketplaces. A higher price increases the absolute size of the commission (even if it is a low percentage), which raises the savings if the buyer and seller bypass the platform (Edelman and Hu, 2016). We believe that cargo delivery services are vulnerable to disintermediation due to the potential higher savings in commissions[10]. Typically, on-demand logistics platforms match drivers (supply) and customers (demand) for delivery requests from point A to point B:

- Driver is the individual that transports goods or passengers in exchange for a payment.
- Customer is the individual or legal entity who enters into a contract of carriage with a driver and pays for a delivery with pre-specified origin and destination.

2.1 Our Application: On-demand Logistics Platform

The setting is a mobile app for on-demand cargo delivery services (like Uber for trucks and cargo vans). The company[11] focuses on intra-city delivery in China and serves 363 cities throughout the whole country. The startup[12] has a valuation of $10 billion. The platform made more than one million matches every day in 2021 by connecting over 7.6 million monthly active customers with 600,000 monthly active drivers with their own vehicles.

Figure 1 illustrates the layout of the app and a transaction from beginning to end. An order starts with the customer requesting a service on the platform. The customer picks

[10] On-demand cargo delivery services features high revenue per job because the delivery involves a longer trip and a larger shipment in general. Appendix A about the industry and gig economy.

[11] The company also operates in 22 international markets across Asia, Latin America, and North America. In the United States, they launch services in Dallas, Houston, and Chicago.

[12] According to the IPO prospectus of a competitor, the market share of our focal platform is 54.7% in 2020, which is ten times the size of the second largest platform with a share of 5.5% in mainland China.

the size of the vehicle, chooses the pick-up time, and sets the origin and destination for the delivery. The platform will assign an available driver to the job based on the driver's distance to the pick-up location. The driver is advised by the platform to call or text the customer on the app to communicate delivery details. The customer can track the driver's location and get updates on their order status in real time. The job can be canceled by the customer without any penalty. The job will be marked as complete after the delivery.

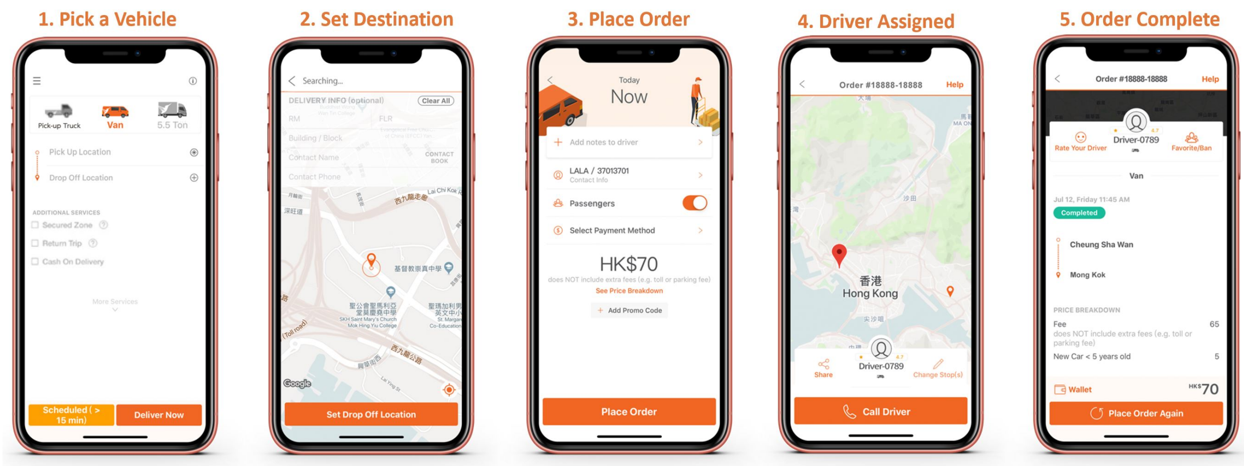


Figure 1: The Order Transaction Flow for Cargo Delivery in the App

In 2019, the platform launched a 15% commission fee in different cities at different times. For example, the platform introduced the fee at Beijing in May and at Shanghai in July. This previously non-existent charge motivates drivers and customers to cancel jobs and coordinate offline to avoid the fee. We observe preliminary evidence on leakage right after the policy change: the average cancellation rates go up from 23.57% to 28.54%, or about 5%, after the drivers are charged commission. The executives of the firm were concerned about leakage[13] upon the launch of commission fee[14]. They were not sure how many of these cancellations could be attributed to leakage. The firm wants to investigate its extent and develop technology to detect and deter disintermediated transactions.

This paper uses data in 2019 before the COVID-19 pandemic. During the staggered

[13] Besides the revenue losses, disintermediation brings reputational risk to the platform because the absence of tracking and review systems might result in poor service quality or safety issues.

[14] The firm wants to experiment with a revenue-sharing agreement with drivers to ride the growth in the volume and value of transactions as more customers join.

rollout of a commission fee, the platform had 4 million registered drivers (300,000 monthly active drivers) and 28 million registered users (4 million monthly active customers) in China. In our data, the platform provided 400,000 to 500,000 matches on a daily basis in 2019, but more than 20% of them were canceled.

Pricing Scheme (The Two-Part Tariff)

Our on-demand logistics platform uses a subscription-based model. To improve the margin, the platform has implemented a commission fee, which is commonly observed in other online marketplaces[15]. Cities that launched a commission have a pre-intervention period (without commission) and a post-intervention period (with commission).

The platform is free for customers to use. However, drivers choose from a menu (see Table 1) to use the free service as Non-VIPs or pay for the premium service as Super-VIPs:

Table 1: The Price Menu for Driver Participants

Tier	Max Daily Jobs	Monthly Membership	Commission Rate	
			Pre-Intervention	Post-Intervention
Non-VIP	2	¥0 ($0)	0%	15%
Super-VIP	∞	¥399 ∼¥799 ($60 ∼ $120)	0%	0%

For all cities, no drivers paid any commission fees from 2013 to 2018 in Policy V1.0. Non-VIPs were free to use the platform, but they could only take up to two jobs per day. If drivers wanted more than two jobs in a day, they would need to pay a subscription fee to become Super-VIPs. In other words, drivers could pay a monthly membership fee upfront, which varies from ¥399 to ¥799 in different cities, to get unlimited job assignments.

In 2019, the platform gradually rolled out Policy V2.0 in a random set of cities, which features a 15% commission rate on Non-VIP drivers. Super-VIP drivers who paid the membership fee could still enjoy unlimited jobs per day without paying any commission. The job assignment was independent of the membership tier. Drivers had equal opportunities to get jobs based on their distance to the pick-up location regardless of being a Super-VIP or not.

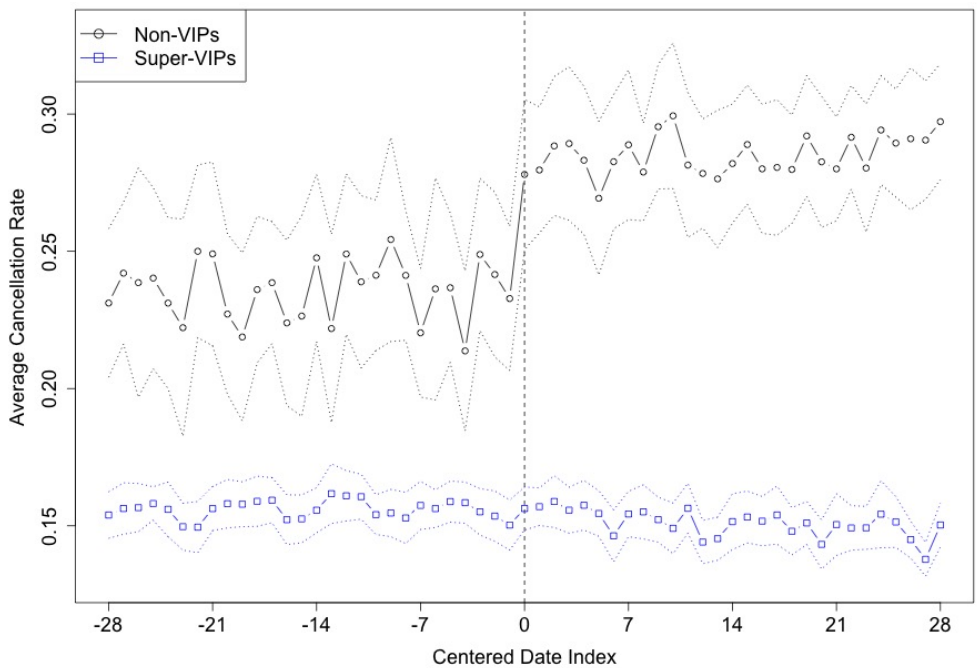
[15] Commissions usually vary between 15-25% for service marketplaces like Uber or Airbnb.

2.2 Preliminary Evidence of Leakage

Between April 20, 2019 and August 31, 2019, the platform randomly launched the 15% commission to 33 cities, contributing 47.9% of matches in the 144 cities with local operational teams[16]. It is a staggered rollout design in different cities at different times.

Figure 2 provides preliminary evidence of leakage. For the 33 treated cities, the average cancellation rates went up from 23.57% to 28.54%, or about 5%, for Non-VIPs who were charged commission after the new policy. There were no changes for Super-VIPs.

Figure 2: The Average Cancellation Rate for 33 Treated Cities



Note: Cancellation rates are averaged across the 33 cities by centering their time series at the launch date of commission. Only Non-VIPs are charged for the 15% commission fee. Super-VIPs have fee waivers. The grey dotted lines are confidence intervals of the average cancellation rates.

The simple pre-post comparison only makes sense when we assume a stable leakage rate without any trending or structural changes other than the commission launch. To provide better suggestive evidence, we implement the synthetic control method (SCM[17]) to evaluate the 33 treated cities. We find that the cancellation rates increased by about 5.17% on average

[16] Although the customers can request service on the app in more than 350 cities in China, only 144 cities have local branches with operational staff that verify vehicles and manage driver relationships.

[17] Appendix C.1 uses Beijing as an example to illustrate how the SCM estimates treatment effects

after charging the fee, using the counterfactual cancellation rates for each treated city that would have occurred had the city not charged a commission fee. The statistical distribution of the SCM estimates are reported in Figure 14 and Table 6 in Appendix C.2.

The preliminary evidence of leakage motivates us to further investigate the problem. However, the cancellation is not a direct measure of disintermediation, and the changes in cancellation rates cannot be fully attributed to leakage without strong assumptions on how the commission fee played a role in our application. In short, we do not know if a given transaction was disintermediated or not. To obtain a direct measure, we combine detailed transaction-level data with geolocation information to trace disintermediation at the individual level. After seeing the 5.17% increase in cancellation rates, the platform had strong motivation to invest in detection technology that utilized the GPS track points and phone conversations collected via its app to flag disintermediated transactions.

3 Leakage Detection and Description

Our unique data[18] come from the on-demand logistics platform. We use geolocation data to check whether a transaction was disintermediated by tracing the GPS footprints of an assigned driver, and use the transaction data to study how leakage responds to platform incentives such as driver-side fees and customer-side subsidies. This section will describe the construction of our leakage dataset and provide descriptive statistics on the key variables.

Geolocation data contain raw GPS track points that record the drivers' location (longitude and latitude) with timestamps. To leverage the staggered rollout design of commission in 2019, we recover the geolocation data in the system to generate labels for disintermediated transactions. Geolocation data are only available between April 20, 2019 and August 31, 2019. During these 137 days, the platform implemented the commission in a random 33 cities, which contributed 47.9% of matches in 144 cities with local operational teams. Since

[18]To maintain privacy, no data contained personally identifiable information that could identify the consumers (e.g., customers) or service providers (e.g., drivers) on the anonymous platform.

the GPS detection algorithm was not available in 2019, there were no penalties (e.g., temporary suspension or permanent ban) on drivers associated with disintermediated transactions because the company did not have ways to verify off-platform transactions.

Transaction data contain delivery jobs that match customers and drivers. Each job has a quoted price, determined by the trip's distance and the requested vehicle's size. The platform only charged commission on Non-VIP drivers in the post-intervention period. Customers received subsidies from marketing experiments or targeting for purposes other than managing leakage. Jobs were characterized by payment methods, item types, time of service, number of passengers, etc. The majority of the jobs were immediate on-demand requests (90%) within the next 15 minutes that requires a small cargo van (57%) for an intra-city delivery (85%) .

We use the geolocation and transaction data to compile the leakage dataset for all 144 cities. The key dependent variable is whether a transaction was disintermediated or not (see Section 3.1 for the variable construction). We will report the extent of disintermediation in Section 3.2. Other variables include drivers' VIP status, commission rates (e.g., 0% vs. 15%), customer coupons, and quoted prices. The data also contains the characteristics of job requests (e.g., on-demand vs. scheduled delivery, escrow vs. cash payment, furniture vs. non-furniture, number of passengers). We provide descriptive statistics on the relationship between disintermediation and these transaction-specific variables in Section 3.3.

3.1 Detection Algorithm

We match the drivers' geolocations with canceled jobs to flag disintermediated transactions:

> The detection algorithm checks whether the driver has GPS track points within the radius of the origin and destination during the time window of a canceled trip, conditional on the customer not having other completed trips that share similar characteristics.

The detection algorithm[19] uses information from both the supply and demand sides. Drivers typically leave a trace in space and time with the following timeline:

[19] We can't disclose the value of radius and time bounds to maintain the proprietary nature of algorithms.

1. The customer initiates a job request and specifies:
    - an origin (pick-up location) and a destination (drop-off location)
    - the time of service (deliver now or scheduled)
2. A driver accepts the job assignment.
3. Side communication occurs (e.g., app, phone calls, or in-person meetings).
4. Either the driver or the customer cancels the job on the platform.
5. The driver visits the origin around the time of service.
6. The driver visits the destination sometime later.

To the best of our knowledge, drivers had little incentive to hide their location. Drivers wanted to share their location to get a job assignment[20]. There were no penalties for disintermediation in 2019, so drivers wouldn't turn off their GPS because of relational incentives. The GPS data used in the detection algorithm were uploaded automatically when a driver uses the mobile app or had the app service running in the background on their phones. The app obtained drivers' consent for using the geolocation data for business operations.

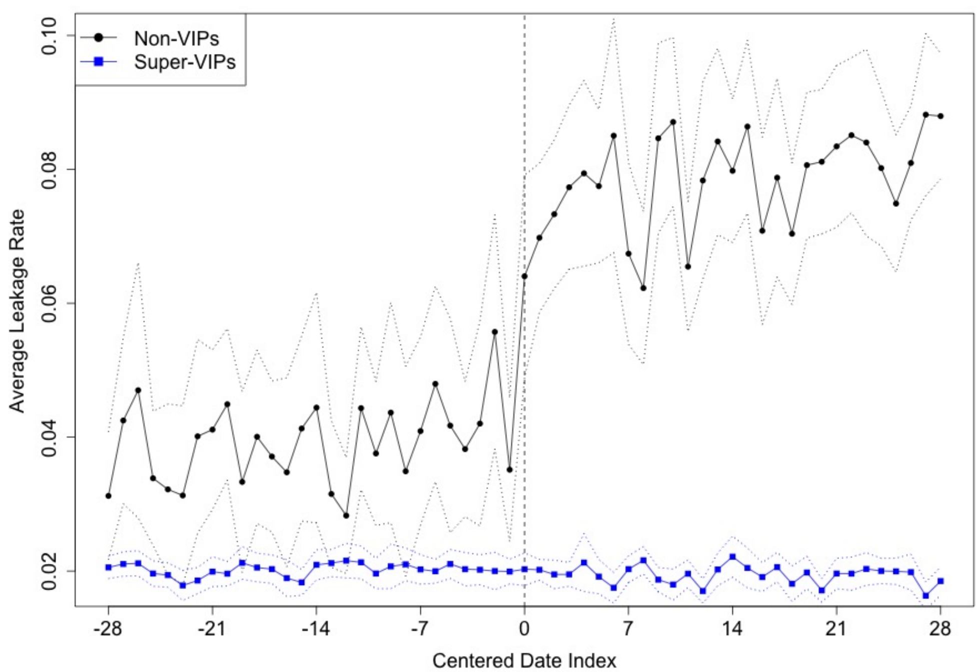
The GPS detection algorithm can achieve both high recall and precision (see Appendix B for more details). Our approach is a direct measure of disintermediation. It does not rely on complicated text analysis and other signals that are indirect measure. According to Xie et al. (2022), platforms would underestimate the scale of leakage if they only use keyword matching to create the measure of intention to disintermediate (Airbnb, 2022b; eBay, 2022b; Gu and Zhu, 2021). Xie et al. (2022) also explore the Bidirectional Encoder Representations from Transformers (BERT), the state-of-the-art model in natural language processing. Results show that the BERT solution is also dominated by the GPS detection. Text mining is not reliable, perhaps because buyers and sellers can encrypt their contact exchanges (e.g., add

[20]The platform assigns a new job to drivers based on their distance to the pickup location; hence, drivers have incentives to keep the platform informed about where they are.

random words between phone numbers) or have their conversations outside the platform (e.g., chat in person or connect on social networks). In our context, drivers face a higher cost[21] to cover up their actions than to hide their intention with words.

We will use the labels created by the GPS detection algorithm for the remaining part of the paper. Xie et al. (2022) show that this approach to identifying disintermediation performs well when evaluated against the human labels as ground truth. As far as we are aware, this study is one of the few, if not only, studies to use such a direct measure of leakage.

Figure 3: The Detected Leakage Rate for 33 Treated Cities



Note: Leakage rates are averaged across the 33 cities by centering their time series at the launch date of commission. Only Non-VIPs are charged for the 15% commission fee. Super-VIPs have fee waivers. The grey and blue dotted lines are confidence intervals of the leakage rates.

3.2 The Extent of Detected Leakage

We checked whether the driver passed the origin and destination of a previously canceled delivery. The assumption is that, if a job was taken offline, the driver would still visit the origin and destination within the previously agreed upon time window for the requested trip.

[21]Drivers who shut down the GPS service would not get any job assignment.

Figure 3 shows the average percentage of disintermediated transactions across the 33 cities by centering their time series at the launch date. The basic impact of commission launch on leakage is evident without any modeling assumptions or control variables. In contrast, we do not see visible changes in the average leakage rate for Super-VIP drivers.

Table 2 reports both the average leakage rates and cancellation rates in 33 cities. The mean pre- and post-intervention leakage rates for Non-VIPs were 3.92% and 7.87%, respectively. Cancellation rates provide an alternative metric for robustness checks. For each city, we summarize the daily percentage of both disintermediated and canceled transactions for 28 days before and 28 days after the launch of commission. Both metrics went up for Non-VIPs but not for Super-VIPs. The increase in average leakage rate for Non-VIP drivers is prominent: the leakage rate increased by 100.7%, or by 3.95 percentage points.

Interestingly, Non-VIP (Super-VIP) drivers disintermediated 3.92% (2.02%) of transactions in the pre-intervention period when they were not charged any commission. These disintermediated transactions were not related to commission fees. For example, drivers and customers might want to change the time of service and shipping addresses. They might also want to deviate from the quoted price by the platform. The delivery contract was thus modified without involving the platform as a third-party. Some Non-VIP drivers might disintermediate to escape the daily caps of jobs they can perform. This behavior might explain the difference in likelihood of disintermediation between Non-VIPs and Super-VIPs.

Table 2: Changes in Leakage and Cancellation after Charging Commission

	Metric	Pre-intervention (28 days before)	Post-intervention (28 days after)	Change in p.p.	Incremental Percentage
Non-VIP	% Detected	3.92%	7.87%	+3.95$p.p.$	+100.7%
	% Canceled	23.57%	28.54%	+4.97$p.p.$	+21.09%
Super-VIP	% Detected	2.02%	1.94%	-0.07$p.p.$	-3.47%
	% Canceled	15.59%	15.09%	-0.50$p.p.$	-3.21%

Note: *p.p.* stands for percentage points, the absolute difference of two percentages.

Suppose drivers were the same, albeit not likely, across the two VIP statuses. The numbers suggest that about 2.02% of transactions were taken offline due to the change of contract (e.g., time, location, price), 1.9% of transactions were disintermediated because drivers wanted more than two jobs in a day without paying the membership fee, and 3.95% of transactions were leaked to avoid the per-transaction commission fee.

The simple pre-post analysis[22] at the city level gives us model-free evidence on the potential effect of commission launch on our direct measure of leakage. Such analysis assumes no structural changes other than the commission launch in these 33 treated cities. To alleviate the concerns on changes in time or changes in driver composition, we conduct two extra analyses: (1) Appendix C.1 documents a case study of Beijing and its neighboring cities using the city-level synthetic control methods (SCM); (2) Appendix D.1 uses the driver-level difference-in-difference (DiD) regressions. We discuss the pros and cons of these two approaches in Appendix D.2 regarding whether the additional analyses is suggestive or causal.

3.3 Descriptive Statistics

We construct the leakage dataset based on a random sample of anonymous drivers with at least one job assignment within the 137 days, tracking all their assigned jobs, cancellations, and VIP status. We randomly draw 1971 drivers[23] from the 144 cities and find that they interacted with 239,057 customers, generating a total of 269,921 matches for our study.

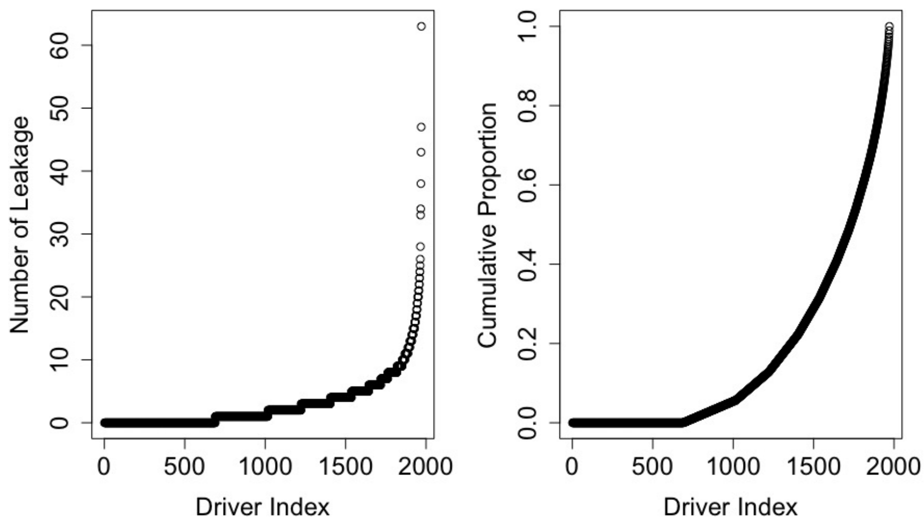
Using the GPS detection algorithm, we find that 2/3 of them (see Figure 4) were involved in at least one disintermediated transaction over the 137 days. An average driver was active for 47/137 days, responding to 2.53 jobs with 0.55 being canceled per active day.

The dataset also contains the transaction price, driver-side commission fees, and customer-side subsidies. Jobs were characterized by payment types, item types, client types, distance, the time of service, the number of passengers, and the size of the vehicle requested.

[22]Figure 14 and Table 6 report the statistical distribution of differences across cities (Appendix C.2).

[23]We sample individuals across geographical regions to obtain a sample with good representation of heterogeneity in driver types making decisions in different market conditions.

Figure 4: Frequency of Disintermediation by 1971 Drivers



Note: The left figure lists the number of disintermediated transactions conducted by each driver in a total of 137 days. The right figure shows the total share of disintermediated transactions cumulated from drivers who had zero offline transactions to the driver who had 63 offline transactions.

Monetary Incentives and Disincentives

This section documents how leakage changed with platform incentives (e.g., driver fees and customer coupons). During the 137 days of the staggered rollout, 4.2% of transactions were charged with a 15% commission on drivers. The average commission fee for these transactions was ¥16.5, which was approximately the minimum hourly wage ($2∼$4) in China.

The commission fee was a previously non-existent charge, which created an incentive conflict between the platform and the driver. The potential savings in commission fees are monetary incentives for drivers to take the transaction offline.

Figure 5 shows that more leakage happened in the transactions with a 15% commission than in transactions without commission (e.g., the gap between the black and blue lines). Moreover, the natural variation in quoted price demonstrates that leakage was more likely to occur for high-value jobs than for low-value jobs when there was a commission fee. In other words, increasing the quoted price, even if the commission rate were unchanged, might increase leakage. The commission shock and natural variation in the quote price allow us to identify the effects of commission on leakage.

Figure 5: Leakage Rate by Price

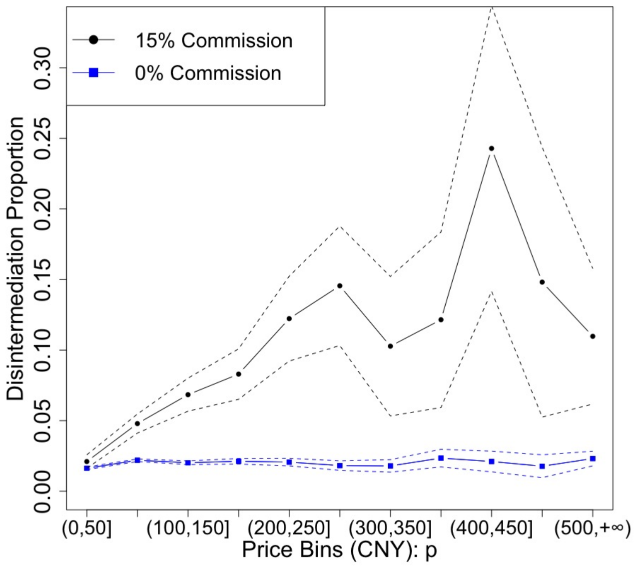


Figure 6: Histogram of Price

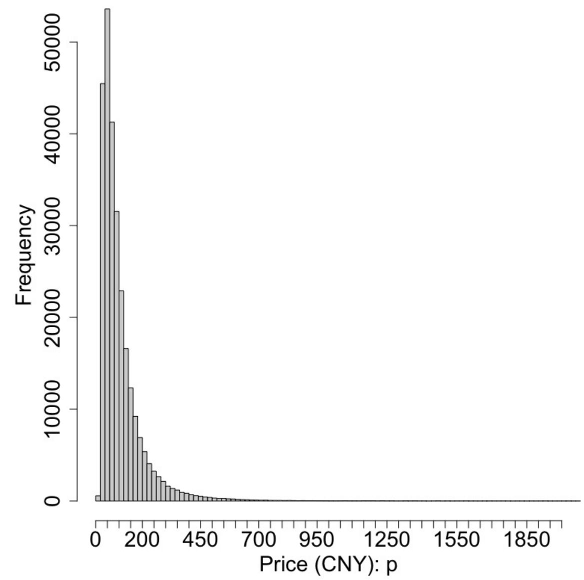


Figure 7: Leakage Rate by Coupon

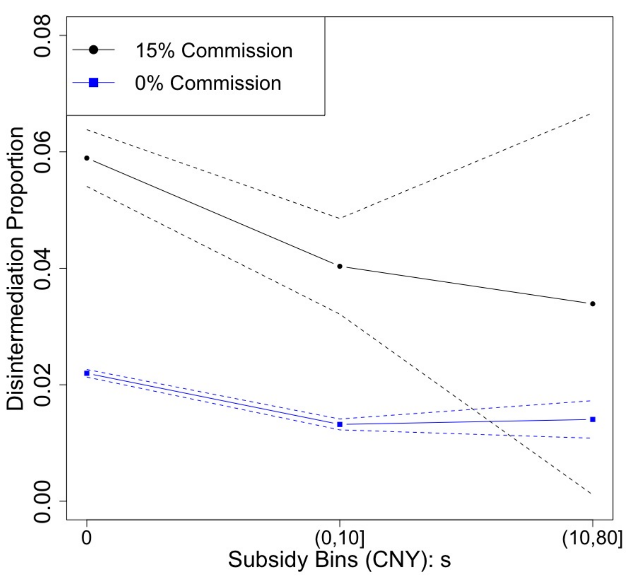
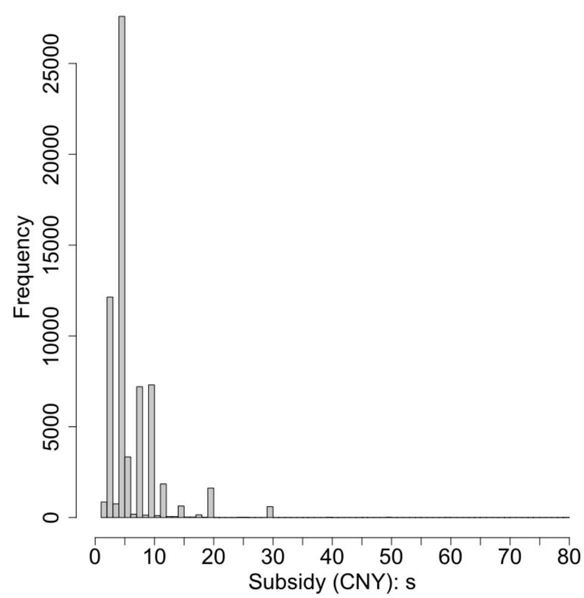


Figure 8: Histogram of Coupon



Note: The average leakage rates are computed on the bins of quoted price or subsidy, conditional on whether the transaction was charged a 15% commission. The dotted lines are confidence intervals.

Figure 7 shows that less leakage happened for subsidized transactions. While commission fees created incentive conflicts, coupons might retain customers. In the sample, we have 23.9% of transactions subsidized on the customer side for promotion or pricing experiments. Most coupons offered a ¥5 discount. The average subsidy for these transactions was ¥6.5, which was approximately $1. These customer subsidies were valid disincentive shocks to leakage because coupons were not distributed to customers for leakage reduction. They were part of marketing experiments to vary quoted prices for customer acquisition.

Transaction-specific Characteristics

The platform specialized in matching customers with drivers for intra-city deliveries (85%) and some occasional inter-city deliveries (15%). Shipments included both merchandise delivery (90%) and furniture delivery (10%) for clients who needed help with moving. More than half of the jobs demanded small vans (57%), followed by small trucks (20%). There were fewer requests for medium vans (13.5%) and medium trucks (9.5%); however, drivers who owned larger vehicles were eligible to take jobs with a smaller load.

Table 3 demonstrates the descriptive statistics of transaction-specific characteristics. For example, customers specified the payment method for their delivery requests. In our sample, 78.2% of transactions proceeded with the escrow payment system provided by the platform ($is\_cash = 0$) and 21.8% of them used cash payment ($is\_cash = 1$). The percentage of detected disintermediation were 4.3% for cash transactions, which was almost three times as high as the 1.5% observed in transactions via escrow payment. The canceled transactions were 36.1% and 15.1% for cash and non-cash transactions, respectively.

The 24/7 on-demand delivery app makes speedy matching in almost real time. Most jobs were immediate (90%) on-demand requests rather than scheduled (10%) shipment requests. It seems that scheduled jobs were much more likely (29% vs.18.6%) to be canceled than the on-demand jobs, which requested a vehicle within 15 minutes, perhaps due to the change in demand or the availability of drivers. The disintermediated transactions detected by the

Table 3: Descriptive Statistics of Transaction-Specific Characteristics

Type	Variable	Value	Freq	% Detected	% Canceled
Payment	is_cash	0	78.2%	1.5%	15.1%
		1	21.8%	4.3%	36.1%
Item	is_furniture	0	89.5%	2.4%	21.9%
		1	10.5%	0.1%	1.0%
Time	is_scheduled	0	89.1%	2.1%	18.6%
		1	10.9%	2.5%	29.0%
Distance	is_intercity	0	85.5%	2.2%	18.7%
		1	14.5%	2.0%	25.7%
Client	is_bus_ep	0	97.4%	2.2%	19.8%
		1	2.6%	1.4%	16.7%
Custodian	passenger	0	85.5%	2.2%	20.1%
		1	8.6%	1.9%	16.7%
		2	5.9%	1.9%	18.4%
Capacity	vehicle	Van_S	57.0%	1.9%	16.6%
		Van_M	13.5%	2.1%	21.3%
		Truck_S	20.0%	2.4%	23.2%
		Truck_M	9.5%	3.0%	28.8%

GPS algorithm is 0.4 percentage point, or 19%, higher for scheduled jobs.

Most customers were individual consumers or small business owners. Enterprise customers contributed 2.6% of transactions with lower probability of disintermediation. Customers could send up to two passengers[24] to accompany the goods in transit as custodians.

In summary, Section 3.3 describes the important aspects of incentive changes with particular emphasis on the variation in the data for identification strategies of our structural model estimates. With the policy changes documented in Section 2.1, we have three main sources of variation in incentives from the commission fee: (i) the introduction of commission to drivers in the treatment cities, (ii) different cities within the country adopted the commission rate at different points in time, and (iii) commission fees differed across the transactions as quoted prices vary by the distance and vehicle types. Subsidy on the customer side provides another source of variation because coupons were not issued to reduce leakage.

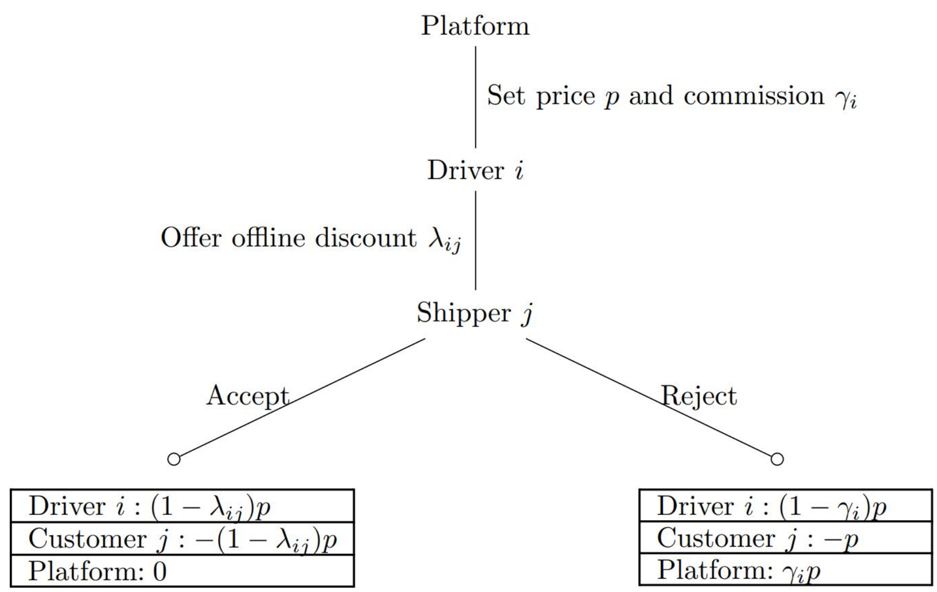
[24]The Road Traffic Safety of China prohibits the usage of cargo vans and trucks for rideshare service.

4 Model

This section develops the model to investigate monetary incentives (Rochet and Tirole, 2006; Weyl, 2010), transaction costs (Coase, 1937; Williamson, 1987; Hagiu and Wright, 2022), and bargaining (Nash, 1950; Rubinstein, 1982) that affects disintermediation. We assume complete information within the driver-customer pair, but the unobservables prevent platforms from perfectly predicting the offline coordination before making a match.

The platform matches i_{th} driver with j_{th} customer who initiates a job request with a quoted price p. Figure 9 demonstrates the monetary transfers between the driver and customer when the transaction is conducted either on the platform or outside the platform:

Figure 9: The Monetary Transfers



Note: Driver i offers an offline discount to propose leakage. Customer j can accept the driver discount λ_{ij} and go off the platform, or reject the proposal and stay on the platform where the driver is responsible to pay the commission fee $\gamma_i p$ to the platform.

The driver pays commission fee $\gamma_i p$ to complete the transaction on the platform. One can consider $\gamma_i p$ as the bid-ask spread set by the platform. The bid-ask spread depends on the elasticity of demand and supply, and transaction costs including moral hazard (Spulber, 2019). We will formally set up the utility functions that take into account the price elasticities and transaction costs in Section 4.1.

The driver can set an offline price $(1-\lambda_{ij})p$ for direct deals by offering the customer a personalized discount λ_{ij}. Survey-based research (Bellotti et al., 2017) documents the existence of offline discounts in ride-sharing services (see Figure 11 in Appendix A). We also have anecdotal evidence of offline discounts in our on-demand cargo delivery platform.

4.1 Utility Functions

To be consistent with the data, we index each match by t that helps us to locate the driver-customer pair using $ij(t)$. The quoted price p_t for the transaction t are non-negative. The platform set a driver-side commission rate $\gamma_{i(t)} \in \{0\%, 15\%\}$ for a subset of drivers.

We denote $\Pi_i^{L_t}(\gamma_{i(t)}, p_t)$ as the driver's utility and $U_j^{L_t}(p_t)$ as the customer's utility, which share the superscript $L_t \in \{1, 0\}$ as the leakage outcome of a joint decision. We make the parametric assumption that the utility functions have the following linear forms:

1. In an on-platform transaction, the payoff functions for $L_t = 0$ are:

$$\begin{aligned} \Pi^0_{i(t)} &= \beta_i \cdot (1-\gamma_{i(t)})p_t \\ U^0_{j(t)} &= u_{j(t)} - \beta_j \cdot (p_t - s_{j(t)}) \end{aligned} \tag{1}$$

   where β_i and β_j are parameters[25] that reflect the heterogeneous marginal utility for money (Dworczak et al., 2021). The platform offers marketing incentives $s_{j(t)}$ to customers (e.g., coupon) to make it cheaper to use the platform services. Customer obtain the baseline utility $u_{j(t)}$ if the driver fulfill the job t.

2. In an off-platform transaction, the payoff functions for $L_t = 1$ are:

$$\begin{aligned} \Pi^1_{i(t)} &= \beta_i \cdot (1-\lambda_{ij(t)})p_t - h_{i(t)} \\ U^1_{j(t)} &= u_{j(t)} - \beta_j \cdot (1-\lambda_{ij(t)})p_t - h_{j(t)} \end{aligned} \tag{2}$$

[25] The utility changes given a unit change in the money the driver receives or the customer pays.

where $h_{i(t)}$ and $h_{j(t)}$ are the relative hassle for the driver[26] and customer[27] to transact outside the platform, respectively. Without the platform governance, driver i can adjust the quoted price p_t with customer j via a contractible term $\lambda_{ij(t)}$ by bargaining (Nash, 1950; Rubinstein, 1982). When $\lambda_{ij(t)} \in [0,1]$, the driver offers a discount[28] to the customer. In other words, there exists an offline trading price at $(1-\lambda_{ij(t)})p_t$ for both sides to agree on the cancellation and then coordinate offline.

4.2 Discrete Choice Models

Driver i is more likely to prefer leakage if $\Pi^0_{i(t)} < \Pi^1_{i(t)}$. Customer j may want leakage if $U^0_{j(t)} < U^1_{j(t)}$. As the platform and researchers, we observe three outcomes:

1. When they collude on leakage, they receive $\Pi^1_{i(t)}$ and $U^1_{j(t)}$.
2. When they transact on the platform, they receive $\Pi^0_{i(t)}$ and $U^0_{j(t)}$.
3. When they cancel the job without collusion on an offline transaction, they receive π^0_i and u^0_j as their outside options. We can normalize the outside options to zero when forming the latent utility gain from leakage.

4.2.1 Driver's Decision

For driver i on transaction t, the latent utility gain from leakage is

$$\begin{aligned}
\Delta\pi_{ij(t)} &= \Pi^1_{i(t)} - \Pi^0_{i(t)} \\
&= \left[\beta_i \cdot (1-\lambda_{ij(t)})p_t - h_{i(t)}\right] - \left[\beta_i \cdot (1-\gamma_{i(t)})p_t\right] \\
&= \beta_i \cdot \underbrace{(\gamma_{i(t)} - \lambda_{ij(t)})p_t}_{\text{Offline Savings}} - h_{i(t)}
\end{aligned} \tag{3}$$

[26] Driver's hassle includes the expected efforts on persuasion, communication, and getting payments.

[27] Customer's hassle includes the inconvenience to track the delivery progress, the efforts to get a proof of payment, and the expected cost of dispute settlement if the goods are damaged or stole.

[28] The driver and the customer can bargain and reach an agreement on how to share the commission savings. For example, $\lambda_{ij(t)} = 0.5\gamma_{i(t)}$ indicates a split of the commission in half.

Driver's decision on leakage depends on the latent indirect utility:

$$L_{i(t)} = \begin{cases} 1 & \Delta\pi_{ij(t)} + \epsilon_{ij(t)} \geq 0 \\ 0 & \Delta\pi_{ij(t)} + \epsilon_{ij(t)} < 0 \end{cases} \tag{4}$$

The $\epsilon_{ij(t)}$ is the structural error, which enters into the driver's final decision of leakage. It include traffic conditions and new information that are not observable to the econometrician.

4.2.2 Customer's Decision

Similarly, customers j has the following latent utility gain from leakage:

$$\begin{aligned} \Delta u_{ji(t)} &= U^1_{j(t)} - U^0_{j(t)} \\ &= [u_{j(t)} - \beta_j \cdot (1 - \lambda_{ij(t)})p_t - h_{j(t)}] - [u_{j(t)} - \beta_j \cdot (p_t - s_{j(t)})] \\ &= \beta_j \cdot (\underbrace{\lambda_{ij(t)} p_t}_{\text{Transfer}} - s_{j(t)}) - h_{j(t)} \end{aligned} \tag{5}$$

Whether to leak is a binary choice depending on the latent indirect utility:

$$L_{j(t)} = \begin{cases} 1 & \Delta u_{ji(t)} + \epsilon_{ji(t)} \geq 0 \\ 0 & \Delta u_{ji(t)} + \epsilon_{ji(t)} < 0 \end{cases} \tag{6}$$

The $\epsilon_{ji(t)}$ is the structural error that enters into the customer's decision of leakage.

4.3 The Joint Decision of Disintermediation

Leakage happens if and only if $L_{i(t)} = L_{j(t)} = 1$ in which both the driver and customer have non-negative indirect utility gain from offline coordination. The platform does not observe the $\epsilon_{ij(t)}$ and $\epsilon_{ji(t)}$, and thus whether leakage occurs or not follows a probabilistic distribution:

$$Pr[L_t = 1] = Pr[\Delta\pi_{ij(t)} + \epsilon_{ij(t)} \geq 0, \Delta u_{ji(t)} + \epsilon_{ji(t)} \geq 0] \tag{7}$$

Bargaining (Nash, 1950; Rubinstein, 1982) happens when two agents can create a surplus together but requires a solution to split the surplus. Assume that the driver-customer pair ij can reach an agreement of $\lambda^*_{ij(t)}$[29] as long as the joint utility gains from leakage is non-negative as shown in the Equation (8). At the equilibrium[30], the $\lambda^*_{ij(t)}$ redistributes the joint surplus $(\Delta\pi_{ij(t)} + \Delta u_{ji(t)})$ between the the driver and the customer to ensure that both individual utility gains ($\Delta\pi^*_{ij(t)}$ and $\Delta u^*_{ji(t)}$) are non-negative[31] before the idiosyncratic shocks kick in. We can thus rewrite Equation (7) into Equation (8) when such $\lambda^*_{ij(t)}$ exists.

$$
\begin{aligned}
Pr[L_t = 1] &= Pr[\Delta\pi^*_{ij(t)} + \epsilon_{ij(t)} + \Delta u^*_{ji(t)} + \epsilon_{ji(t)} \geq 0] \\
&= Pr[\beta_i \cdot (\gamma_{i(t)} - \lambda^*_{ij(t)})p_t - h_{i(t)} + \epsilon_{ij(t)} \\
&\qquad + \beta_j \cdot (\lambda^*_{ij(t)}p_t - s_{j(t)}) - h_{j(t)} + \epsilon_{ji(t)} \geq 0] \\
&= F_\epsilon \left[\beta_i \cdot \gamma_{i(t)}p_t - (\beta_i - \beta_j)\lambda^*_{ij(t)} \cdot p_t - \beta_j \cdot s_{j(t)} - h_{i(t)} - h_{j(t)}\right]
\end{aligned}
\tag{8}
$$

The $\epsilon_{ij(t)}$ and $\epsilon_{ji(t)}$ are structural error terms that enter into the decisions of driver and customer, respectively. Structural errors (McFadden and Train, 2000) are typically introduced to account for idiosyncratic shocks (e.g., new information) and unobservables (e.g., traffic conditions). In our application, quoted prices are orthogonal to the error terms (e.g., the platform does not use dynamic pricing[32] based on real-time market or traffic conditions).

5 Estimation and Results

A proper econometric setting requires that we carefully distinguish what the econometrician can observe from unobserved heterogeneity, which only the driver-customer pair observes in their match. The econometrician cannot observe all the determinants of the pre-transfer utilities of driver i and customer j (Galichon and Salanié, 2021). Specifically, we don't know

[29] I use Nash Bargaining to microfound $\lambda_{ij(t)}$ in Appendix G that describes the the agreement process

[30] The solutions of $\lambda^*_{ij(t)}$ can exist for the binding Individual Rationality (IR) constraints.

[31] Drivers and customers can agree when the $\lambda^*_{ij(t)}$ makes one party indifferent between transacting offline or online. The bargaining power determines whether the driver or customer's IR constraint is binding.

[32] The on-demand cargo delivery service platform determine the pricing rule at the city level, which is a linear function of the trip distance conditional on the requested vehicle size

the exact individual β_i, β_j, $h_i(t)$, and $h_j(t)$ in additional to the structural error terms.

What we know are the observed characteristics: types $d \in D$ for drivers (service providers) and $s \in S$ for shippers[33](customers). These types are observed by everyone as well as the econometrician. In other words, richer data converts unobserved heterogeneity into types.

In the following, we denote $\beta_i = \beta_d$ and $h_i(t) = h_d(t)$ if driver i is of type d, and $\beta_j = \beta_s$ and $h_j(t) = h_s(t)$ if customer j is of type s. Equation 8 becomes Equation 9:

$$Pr[L_t = 1] = F_\epsilon \underbrace{\left[\beta_d \cdot \gamma_{i(t)} p_t - (\beta_d - \beta_s)\lambda^*_{ds} \cdot p_t - \beta_s \cdot s_{j(t)} - (h_{d(t)} + h_{s(t)})\right]}_{x'_{ij(t)}\theta} \tag{9}$$

$$\text{where} \quad d \in D, s \in S$$

5.1 Identification Strategies

Variation in the driver-side commission $\gamma_{i(t)} p_t$ (Figure 5) identifies β_d, the drivers' marginal utility for money. Variation in the customer-side subsidy $s_{j(t)}$ (Figure 7) identifies β_s, customers' marginal utility for money. Since we observe natural variation in the quoted price p_t (Figure 6), we can back out the type-specific offline discount $\lambda_{ds(t)}$ from the price coefficient $-(\beta_d - \beta_s)\lambda^*_{ds(t)}$ given that β_d and β_s are identified.

The remaining part is the offline hassle $h_t = (h_{d(t)} + h_{s(t)})$. They are constant terms that can be decomposed by types of transactions (see Table 3). The leakage rates were different in jobs for intra-city deliveries and inter-city deliveries. Requests for furniture delivery were much less likely to be disintermediated than for merchandise delivery. Most jobs were immediate on-demand requests which suffered less from disintermediation than the scheduled delivery requests. While escrow service with electronic payment through the platform was available, about 20% of job requests specified cash as the payment method and had a much higher probability of being disintermediated.

An alternative identification strategy of $h_{d(t)}$ is to utilize the within-individual[34] variation

[33]Customers are called shippers into the industry because they are the individual or legal entity who enters in a contract of carriage with a driver and pays the driver for delivery.

[34]Appendix E provides some examples to illustrate the intuition for identification within an individual.

of commission fee. By introducing driver-fixed effects, I can back out h_i given the intuition that driver i are more likely to be involved in disintermediated transaction when $\gamma_{i(t)}p_t > h_i$ but stay on the platform when $\gamma_{i(t)}p_t < h_i$ after controlling for other confounding variables. The average hassle for drivers is thus the mean of h_i. Observing multiple transactions within a driver enables the identification. In my sample, an average driver was assigned 2.53 orders per active day. About 2/3 of drivers were involved in at least one offline transactions.

The data $x_{ij(t)} = [\gamma_{i(t)}, s_{j(t)}, p_t, x_t, x_{ij}]$ contains the policy shock on commission rates (e.g., 0% vs. 15%), the platform's subsidy to customers, and the trading price of the transaction. Moreover, the data also contains transaction-specific characteristics (e.g., on-demand vs. scheduled delivery, escrow vs. cash payment, furniture vs. non-furniture, number of passengers) and the drivers and customers indices. The goal is to estimate $\theta = [\beta_d, \beta_s, \lambda^*_{ds(t)}, h_{d(t)} + h_{s(t)}]$

5.2 Maximum Likelihood Estimation

Following the long tradition in discrete choice models (McFadden and Train, 2000), we assume that $\epsilon_t = \epsilon_{ij(t)} + \epsilon_{ji(t)}$ are independently and identically distributed (IID) errors[35] drawn from the standard Type I Extreme Value distribution (i.e., Gumbel). We can use maximum likelihood estimation (MLE) to estimate the standard discrete choice model depicted in the Equation (9) with the data $x_{ij(t)} = [\gamma_{i(t)}, s_{j(t)}, p_t, x_t, x_{ij}]$. We recover the vector of parameter $\theta = [\beta_d, \beta_s, \lambda^*_{ds(t)}, h_{d(t)} + h_{s(t)}]$ by maximizing the following log likelihood function:

$$\mathcal{L}(\theta) = \sum_{t=1}^{N} L_t \cdot ln\left[\frac{exp(x'_{ij(t)}\theta)}{1 + exp(x'_{ij(t)}\theta)}\right] + (1 - L_t) \cdot ln\left[\frac{1}{1 + exp(x'_{ij(t)}\theta)}\right] \tag{10}$$

To obtain insights, we will start with estimating a model that assumes homogeneous price sensitivity of driver and customer and a universal $\bar{\lambda}$ across different transaction types. We will then introduce driver fixed-effects and customer fixed-effects to understand how the

[35] Alternatively, we can use partial identification (Manski, 2003; Manski, 2007) to compute bounds that summarize what the data say about the parameters without Type-I error assumption. See Appendix E.

market responds to platform incentives for a subset of active drivers and customers. Lastly, we will introduce the heterogeneity of price sensitivity in different jobs.

5.3 Homogeneous Price Sensitivity

The simplest model to estimate is to assume that there is one type of driver (only one $d \in D$) and one type of customer (only one $s \in S$). In this model, we will only use $x_{ij(t)} = [\gamma_{i(t)}, s_{j(t)}, p_t, x_t]$ to estimate $\theta = [\beta_d, \beta_s, \bar{\lambda}, h_t]$ with the following assumptions:

- Assume homogeneous price sensitivity β_d and β_s for drivers and customers
- Assume linearly separable hassle $h_t = h + x_t^{'}\beta_{h(T)}$ for jobs of type T
- Assume a universal offline discount $\lambda^*_{ds(t)} = \bar{\lambda}$ that drivers and customers agree on in the market to split the surplus from an offline transaction

The MLE of Equation (9) can be estimated using the entire sample of transactions with

$$x_{ij(t)}^{'}\theta = -h - x_t^{'}\beta_{h(T)} + \beta_d \cdot \gamma_{i(t)}p_t - \beta_s \cdot s_{j(t)} - (\beta_d - \beta_s)\bar{\lambda} \cdot p_t \tag{11}$$

where $\gamma_{i(t)}p_t$ is the driver-side commission fee, p_t is the quoted price for the job, and $s_{j(t)}$ is the customer-side subsidy, x_t are transaction-specific covariates.

Results for assuming IID transactions are reported in the Column (1) and Column (2) in Table 4. Using the model that controls for VIP status, we identify $\hat{\beta}_d = 0.195$ and $\hat{\beta}_s = 0.025$ from the variation in $\gamma_{i(t)}p_t$ and $s_{j(t)}$, respectively. Since $\hat{\beta}_d > 0$, leakage increases in driver commission in our sample. However, leakage is insensitive to customer coupons[36]. Although leakage decreases in the subsidy, customer coupons might not be an effective lever.

The above estimates help us to back out the surplus division rule in rational expectation. On average, drivers and customers agree on an offline discount that roughly split the commission savings in half. We obtain this insight by leveraging the differential price sensitivities

[36]The managerial implications of $\beta_d > \beta_s$: it might be more effective to reduce leakage by subsidizing drivers than customers because drivers were in general more price sensitive

Table 4: Homogeneous Price Sensitivity with a Uniform Offline Discount

	Dependent variable: Disintermediation			
	(1)	(2)	(3)	(4)
commission_fee ($\gamma_{i(t)} p_t$)	0.244***	0.195***	0.205***	0.175***
	(0.015)	(0.015)	(0.032)	(0.031)
subsidy_to_customer ($s_{j(t)}$)	-0.024	-0.025	-0.029	-0.027
	(0.049)	(0.049)	(0.051)	(0.051)
transaction_price (p_t)	-0.015***	-0.013***	-0.014***	-0.013***
	(0.002)	(0.002)	(0.003)	(0.002)
is_cash	1.01***	1.02***	1.02***	1.02***
	(0.029)	(0.029)	(0.048)	(0.048)
is_furniture	-3.08***	-3.08***	-3.14***	-3.14***
	(0.173)	(0.173)	(0.175)	(0.175)
is_scheduled	0.458***	0.435***	0.390***	0.389***
	(0.042)	(0.042)	(0.047)	(0.047)
is_intercity	-0.202***	-0.220***	-0.151***	-0.162***
	(0.045)	(0.045)	(0.057)	(0.055)
is_bus_ep	-0.148	-0.170	-0.168	-0.187
	(0.107)	(0.107)	(0.114)	(0.114)
passenger1	0.199***	0.200***	0.165***	0.166***
	(0.051)	(0.051)	(0.054)	(0.054)
passenger2	0.327***	0.334***	0.305***	0.308***
	(0.062)	(0.062)	(0.068)	(0.068)
vehicleTruck_M	0.497***	0.417***	0.372**	0.336*
	(0.047)	(0.047)	(0.186)	(0.184)
vehicleTruck_S	0.261***	0.219***	0.102	0.087
	(0.035)	(0.035)	(0.108)	(0.107)
vehicleVan_M	0.117***	0.113***	0.036	0.030
	(0.042)	(0.042)	(0.070)	(0.070)
vipDriver		-0.473***		-0.546***
		(0.036)		(0.083)
(Intercept)	-4.08***	-3.68***		
	(0.027)	(0.040)		
Fixed-effects: driver_id			✓	✓
Average Discount ($\bar{\lambda}$)	0.068	0.074	0.079	0.085
Observations	269,921	269,911	248,556	248,556
- Unique Drivers	1971	1962	1280	1280
- Unique Users	239,057	239,048	220,920	220,920
Pseudo R^2	0.05059	0.05341	0.10887	0.11021
BIC	53,092.2	52,946.8	64,863.2	64,802.0

of drivers and customers. Given that we identify $\hat{\beta}_d$ and $\hat{\beta}_s$, the coefficient for price, which is $-(\hat{\beta}_d - \hat{\beta}_s)\bar{\lambda} = -0.013$, tells us that $\hat{\bar{\lambda}} = 0.074$.

Leakage would occur when the total commission savings exceed the offline transaction costs. The estimates, $-\hat{h}/(\hat{\beta}_d - \hat{\beta}_s)$, show that the offline costs in transactions with Non-VIPs[37] are about ¥21.62 ($3) for the most common type of transactions: the on-demand and intra-city delivery of goods requested by non-enterprise customers that use digital escrow payment service. The estimated transaction costs are higher than the average commission fee, ¥16.5, the platform receives in our sample.

People who disintermediate should have lower costs to take transactions offline than those who stay on the platform. In our sample, the average offline cost of the $n_1 = 5761$ disintermediated transactions is ¥20.37, calculated by $\frac{1}{n_1}\sum(\hat{h} + x'_{t(L=1)}\hat{\beta_{h(T)}})/(\hat{\beta}_d - \hat{\beta}_s)$. It is smaller than the ¥23.91 estimated for the $n_0 = 264160$ non-disintermediated transactions by $\frac{1}{n_0}\sum(\hat{h} + x'_{t(L=0)}\hat{\beta_{h(T)}})/(\hat{\beta}_d - \hat{\beta}_s)$. These two estimates are consistent with what we expected as offline transaction costs would prevent leakage from happening.

Relative transaction costs can inform us about the value of platform services. For example, transactions are more likely to be disintermediated when customers choose to pay cash instead of using the digital escrow payment service. Only 21.8% of transactions specified cash as the payment method. The relative cost of offline transaction is lower by ¥6 ($1) for these cash-paying jobs. We can interpret this dollar value as how much the driver-customer pairs were implicitly paying for using the platform payment system.

Suggestive evidence shows that less moral hazard or easier communication within the driver-customer pair is associated with higher leakage. Estimates show that the cost of coordination outside the platform is lower when customers have at least one passenger on board with the delivery. The likelihood of disintermediation is higher if customers send two passengers instead of one passenger. In cargo delivery services, most passengers are custodians to accompany the goods and supervise the delivery. It is illegal to use cargo van

[37]The average offline transaction cost of Super-VIPs is ¥2.78 higher than that of Non-VIPs.

and trucks for rideshare service according to the Road Traffic Safety Law in china.

We find that on-demand delivery (ship now) is less likely to be disintermediated while non-urgent (scheduled) delivery is positively associated with leakage. It is interesting to see that intercity delivery has lower leakage than intracity delivery. A larger load size ($Truck_M > Truck_S > Van_M > Van_S$) is associated with a higher leakage rate. Lastly, delivery requests from large enterprise customers (about 3%) or furniture moving (about 10% of total transactions) are associated with a lower probability of disintermediation.

5.3.1 Driver Heterogeneity in Transaction Costs

We now exploit the within-driver variation to back out individual hassle for a subset of active drivers (2/3 of my sample). These 1280 drivers had repeated transactions with at least one disintermediated transactions. The MLE of Equation (9) with driver fixed-effects has

$$x'_{ij(t)}\theta = -h_i - x'_t\beta_{h(T)} + \beta_d \cdot \gamma_{i(t)} p_t - \beta_s \cdot s_{j(t)} - (\beta_d - \beta_s)\bar{\lambda} \cdot p_t \tag{12}$$

Results for are reported in the Column (3) and Column (4) in Table 4. These active drivers seem to be more aggressive in offering an average discount of 8.5%, which is backed out by the model that controls for VIP status and identifies $\hat{\beta}_d = 0.175$ and $\hat{\beta}_s = 0.027$ from variation in the data. Again, we observe that leakage increases in driver commission fees and decreases in customer subsidies. The average relative hassle for offline transactions is -3.519, which translates to ¥23.77 ($3.5) for Non-VIP drivers. The Super-VIP drivers have an additional ¥3.68 ($0.5) cost to transact offline. The insights about platform services from model estimates without driver fixed-effects still hold in this subset of data.

5.3.2 Customer Heterogeneity in Transaction Costs

Similarly to Section 5.3.1, we can exploit the within-customer variation to check how market responded to the commission fee for a subset of active customers. Only 989 out of 239,057 were involved with repeated transactions with our sampled drivers. The MLE of Equation (9) with both driver and customer fixed-effects has

$$x'_{ij(t)}\theta = -h_i - h_j - x'_t\beta_{h_{(T)}} + \beta_d \cdot \gamma_{i(t)}p_t - \beta_s \cdot s_{j(t)} - (\beta_d - \beta_s)\bar{\lambda} \cdot p_t \tag{13}$$

Table 8 in Appendix F show that these experienced customers obtain an offline discount ranged from 16% to 21.6%. Interpretation should be careful given the very small sample of unrepresentative customers and drivers. Estimates of transaction costs in some transaction types do not converge well due to the limited sample size in the category (e.g., furniture movement, number of passengers). Nonetheless, it seems that platforms and drivers might not be benefited from having shrewd customers who had strong bargaining power. These customers might negotiate for a much lower price than the quoted price by the platform.

5.4 Heterogeneous Price Sensitivity

Previously, we assumed homogeneous price sensitivity as we considered only one driver type and one customer type. In this section, we relax the assumption to use observed characteristics of drivers and customers to define types $d \in D$ and $s \in S$ with variables in $x_{ij(t)}$. These types are observed by all market participants as well as the econometrician. Richer data can convert unobserved heterogeneity into types. We include the transaction-specific covariates in x_t to interact with $\gamma_{i(t)}p_t$, $s_{j(t)}$, and p_t to allow heterogeneity in β_d and β_s.

- Assume heterogeneous price sensitivity for drivers $\beta_d = \bar{\beta}_d + x'_{ds}\beta_{d_{(ds)}} + x'_t\beta_{d_{(T)}}$
- Assume heterogeneous price sensitivity for customers $\beta_d = \bar{\beta}_d + x'_{ds}\beta_{d_{(ds)}} + x'_t\beta_{d_{(T)}}$
- Assume linearly separable hassle $h_t = h_i + x'_{ds}\beta_{h_{(ds)}} + x'_t\beta_{h_{(T)}}$
- Assume type specific discount $\lambda^*_{ds(t)} = \lambda + x'_{ds}\beta_{\lambda_{(ds)}} + x'_t\beta_{\lambda_{(T)}}$ that drivers and customers agree on in the market to split the surplus from offline transaction

We can estimate $\theta = [\bar{\beta}_d, \bar{\beta}_s, \beta_{d_{(ds)}}, \beta_{d_{(T)}}, \bar{\beta}_s, \beta_{s_{(ds)}}, \beta_{s_{(T)}}, h_i, \beta_{h_{(ds)}}, \beta_{h_{(T)}}, \lambda, \beta_{\lambda_{(ds)}}, \beta_{\lambda_{(T)}}]$ using the data $x_{ij(t)} = [\gamma_{i(t)}p_t, s_{j(t)}, p_t, x_{ij}, x_t]$ with interactions between the variables. The interaction terms provide fruitful insights into the leakage responses.

For example, interactions between commission fees and transaction-specific characteristics can capture the heterogeneity of drivers' price sensitivities in different contexts. Drivers are less likely to disintermediate a furniture delivery or intercity delivery as the amount of commission fees goes up (see Table 9 in Appendix F). This negative relationship contradicts regular cases where leakage is more likely to happen when the savings in commission fees are higher. Such contradiction indicates that drivers are willing to pay fees to the platform when they handle specific types of jobs, such as moving furniture or driving to another city, perhaps due to their wants for dispute settlement or protection provided by the platform.

Higher commission fees are not always positively associated with a higher likelihood of leakage, specifically, when drivers are involved in cash-paying jobs (negative coefficients for $commission\_fee \times is\_cash$). The risk of losing the payment (negative coefficient for $transaction\_price \times is\_cash$) may outweigh the potential savings in commission fees (positvie coefficient for $commission\_fee$). Drivers may be concerned about being defaulted on high-value jobs when customers want to use cash instead of the escrow payment service on the platform. Adverse selection could explain why drivers are less price sensitive to platform services when customers specify cash payment.

Table 9 of Appendix F reports the model estimates with heterogeneous price sensitivity. We use these rich estimates to conduct the counterfactual analysis in Appendix G to guide the discussion of policy implications in Section 6.3. In the future analysis, we can include more variables of drivers and customers, such as their demographics and RFM (recency, frequency, monetary) statistics in our sample. In the next section, we discuss how we use the model estimates to inform platform design to mitigate leakage in two-sided markets.

6 Implications for Platform Design

Platforms want to neutralize disintermediation to capture the value they have created and retain complete transaction data. To forestall leakage, we focus on ex-ante approaches that better align the incentives between the platform and the driver-customer pairs. We do not consider ex-post punishments in this research, because of the difficulty in verifying which party is at fault in two-sided markets as well as dealing with rebuttals. Preventive measures may provide a better long-term outcome for platforms without losing future revenue from banning accounts or triggering antagonism that reduces platform engagement.

6.1 Marketing Interventions

One way to manage leakage is to provide coupons to get the commission fee just below the offline transaction costs between drivers and customers. The platform can target individuals that are sensitive to fee reduction and personalize the amount of compensation. The optimal targeting rule might differ from targeting on predictive churn. Platforms typically use machine learning to target individuals with the highest probability of leaving, but such retention efforts might be futile when they fail to consider individuals' sensitivity to the intervention (Ascarza, 2018). We might not be able to persuade drivers who will disintermediate anyway. Instead, we can compensate drivers who are more sensitive to commission fees and have a moderate offline transaction costs. This alternative retention strategy might result in a more significant reduction in leakage given the same marketing budget.

In future research, we will conduct a counterfactual analysis to test whether targeting drivers with moderate offline transaction costs and high commission sensitivity is more cost-effective than targeting drivers with the highest risk of leakage. The key objective is to use less money to retain more transactions by converting drivers at the boundary of leakage.

6.2 Monitoring Technology

Platforms can use monitoring technology to increase offline transaction costs or decrease online transaction costs. One straightforward way is to actively listen to conversations between buyers and sellers to block contact exchanges[38]. However, monitoring conversations faces many limitations in on-demand services. It not only triggers privacy concerns[39], but it is also evadable when buyers and sellers can have their conversations outside the platform (e.g., meet in person) or shut down the monitoring devices (e.g., turn off the phone).

A potentially better alternative is to invest in platform technology that can reduce the transaction costs for on-platform transactions. For example, the platform can compensate drivers for installing monitoring devices (e.g., video cameras) in the cargo space (e.g., back truck, trailer, or cargo bed). The customer can request access to the live remote video camera to monitor their goods in transit. This service creates a new incentive for customers to stay on the platform. It may also reduce the need for customers to send someone to accompany the goods, which facilitates leakage[40]. Since 2021, the on-demand cargo delivery service app has been actively promoting Internet of Things (IoT) devices on vehicles to help protect customers' safety and assets. The monitoring devices can also provide evidence to resolve disputes regarding damaged goods. However, no efforts have been made to provide customers with video streams or photo snapshots. The platform can take advantage of the existing technology to reduce the monitoring cost for customers.

6.3 Matching Policy

The platform can strategically match a driver and a customer who have large enough offline transaction costs together as a pair. The variation in commission fee can recover drivers' hassle, and the transaction characteristics of job requests can index the type-specific customers'

[38]Airbnb detects and blocks contact exchanges by replacing emails and phone numbers with "(Hidden by Airbnb)" to stop people from dealing directly with the guest or host.

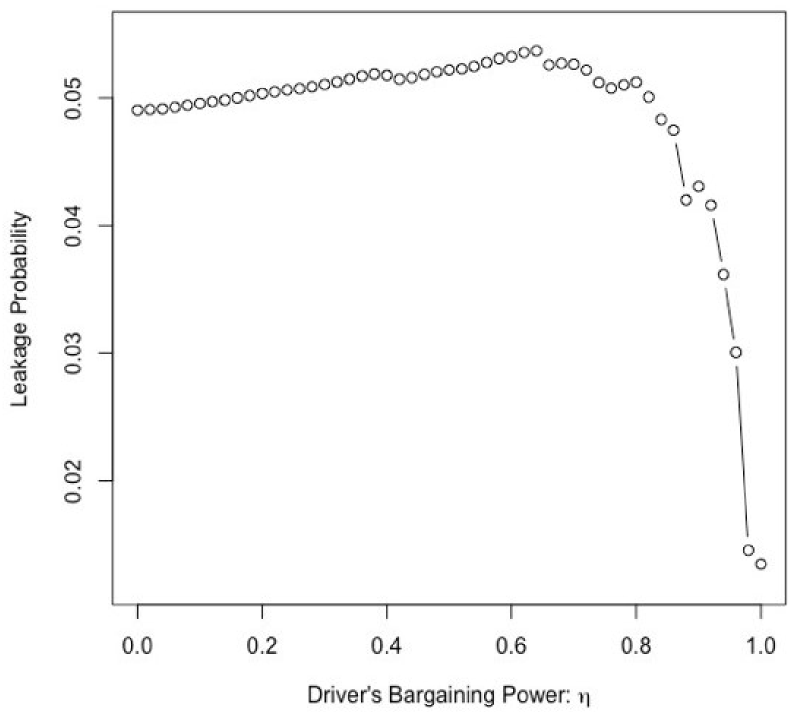
[39]Uber's China counterpart, Didi Chuxing, launched a mandatory audio recording as a safety feature. However, passengers are not buying the feature that trades privacy for safety (Shen, 2019).

[40]Table 5.3.1 shows that more passengers are associated with the lower hassle and higher leakage.

hassle (see Section 5.3.1 for estimates). Given the historical information[41], platforms can set a restriction to make sure that the pair-specific transaction costs for offline transactions are lower than the commission fee they charge for on-platform transactions.

A focus on fast and last-minute matching might also help the platform mitigate leakage. Our counterfactual analysis in Appendix G shows that the likelihood of leakage is higher when drivers have just slightly stronger bargaining power than customers (see Figure 10). The platform would be better off, in terms of less leakage, having full bargaining power on the driver's side. The platform might suffer from more leakage in a market if customers have a stronger bargaining power. The intuition is that customers with a strong bargaining

Figure 10: Counterfactual Leakage Rates from Different Bargaining Power



Note: The likelihood of leakage is highest at $\eta = 0.65$ when drivers have slightly stronger bargaining power than customers (see Appendix G for the micro-founded model that uses Nash Bargaining as the solution concept).

power might actively ask for an offline discount which drivers cannot decline. For example, customers can pay less offline, rather than the full quoted price, when they have patience (e.g.,

[41]For new drivers, the platform can use machine learning to predict hassle based on their characteristics at registration. The screening process of hassle can also inform better driver acquisition or retention.

sufficient time for negotiation) and have outside options (e.g., other competitive drivers). Patience and outside options can empower customers in the bargaining process (Rubinstein, 1982; Backus et al., 2020). Additional descriptive evidence supports the implications of this counterfactual analysis. For example, scheduled jobs have higher leakage, perhaps because the customer has more time to find alternative drivers or negotiate a better deal. However, leakage is lower if the job is scheduled in the early morning for the next day, perhaps due to the lack of supply. Appendix G.3 discusses how market conditions might affect the bargaining power. With the insights from our counterfactual analysis, the platform can experiment with assigning or disclosing drivers to customers at the last minute to mitigate leakage.

7 Discussion

This research was motivated by the finding that the cancellation rates went up after the platform charged a 15% commission rate. Intrigued by this finding, we leverage geolocation data to identify offline transactions that are typically hard to track in online marketplaces. To provide insights into preventive measures, we focus on the critical question of how leakage responds to platform incentives. We estimate the price elasticity of leakage and transaction costs in a structural model by exploiting the quasi-experimental variation in incentives (e.g., driver fees and customer coupons) for leakage. One source of variation comes from the changes in commission fees, which are generated by the staggered rollout of a 15% driver commission across cities and the variation in quoted prices across transactions. Another source of variation comes from the coupons issued to customers for purposes other than reducing leakage, which generates external disincentives for leakage.

We find that, on average, the likelihood of leakage increases as the driver commission fees go up, but it is insensitive to the customer coupons in our sample. At the same time, we find that the result is not uniform and depends on the types of transactions. For example, leakage was more likely to happen when customers specified the payment method as cash.

However, drivers were less likely to disintermediate cash-only jobs with higher quoted prices, even though the potential savings in commission fees were larger. Given the heterogeneity of price sensitivity across transactions, the platform can experiment with differential fees in different transaction contexts to mitigate leakage and improve revenue.

Our novel data and model estimates provide unique insights into preventive measures, which are ex-ante alternatives to ex-post punishments. The platform can prioritize the targeting of drivers with high commission sensitivity and compensate them based on their transaction costs. This targeting rule may be more cost-effective than simply targeting drivers with the highest risk of leakage who might not be sensitive to retention offers. The platform can also leverage the recent rollout of monitoring technology (IoT devices on vehicles) to mitigate leakage. Granting customers access to the live remote video camera in the cargo space to monitor their goods can create standalone value for using the platform. This service may reduce the need for customers to send someone to accompany the goods in transit which could facilitate leakage. Lastly, the platform can experiment with assigning drivers to customers at the last minute and continue advertising on fast matching. A focus on instant delivery can prevent drivers and customers from having sufficient time to negotiate, and reduce the chance for them to find competitive outside options as leverage to bargain for a favorable offline price that is different from the price quoted by the platform.

Given the nature of the economic problem, our empirical framework can inspire analyses of other intermediaries in marketing and financial applications, including, but not limited to, retailers in between manufacturers and consumers, housing agents who mediate homeowners and homebuyers, brokers for private equity and investors, and travel agencies searching airlines and hotels for travelers. Similar to our case, these other applications involve buyers and sellers who make decisions on a daily basis about whether or not to engage in direct sales. We hope this study motivates new platform designs and pricing strategies for intermediaries to mitigate leakage when enforcing minimum advertised pricing (MAP) is impossible.

Limitations and Directions for Future Research

As far as we know, our work is one of the few, if only, studies that uses a direct measure of offline transactions, though the GPS detection is not perfect. The detection algorithm might not recover all disintermediated cases due to data availability and correctness. For example, drivers could turn off their GPS or shut down the phone. Poor cell phone signals might also affect the track-point uploads. However, we believe it was in the driver's best interest to share where they are, because they could not get any job assignment for on-demand delivery without disclosing their locations. Moreover, there were no penalties for disintermediation that would motivate drivers to hide their activities to avoid punishments.

The profit-maximizing commission rate depends on who participates (driver selection). The static model in this paper does not account for the long-run entry and exit decisions. Counterfactual commission rates will affect who stays active on the platform. As a result, the job assignments (matching outcomes) change and affect the leakage rates. One future direction is to develop a dynamic structural model that encompasses the endogeneity in driver selection with the capability of simulating alternative market equilibria.

We believe an important next step is to design the optimal menu of two-part tariffs for two-sided markets. The objective of platforms is to capture as much value as possible rather than to eliminate all disintermediation. Platform businesses can use a set of different combinations of fixed subscription fees and per-transaction fees to screen drivers, whereby the driver's choice reveals their type (e.g., price sensitivity and transaction costs). The mechanism design can extract more surplus from buyer-seller pairs under second-degree price discrimination and may tolerate a certain degree of leakage.

We provide unique insights into preventive measures (e.g., targeting coupons, monitoring technology, and matching algorithms) that reduce leakage. Future research can conduct field experiments to test these ex-ante approaches to mitigate leakage. We hope this study provides initial guidelines for new platform design and motivates more preventive measures in a world full of punishments (Uber, 2022; Airbnb, 2022a; eBay, 2022a).

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program". *Journal of the American Statistical Association* 105.490, 493–505.

Airbnb (2022a). *Airbnb's Off-Platform Policy*. URL: https://www.airbnb.com/help/article/2799/airbnbs-offplatform-policy.

— (2022b). *Why We Review Messages*. URL: https://www.airbnb.com/help/article/1121/why-we-review-messages-on-airbnb.

Ascarza, Eva (2018). "Retention futility: Targeting high-risk customers might be ineffective". *Journal of Marketing Research* 55.1, 80–98.

Athey, Susan and Guido W Imbens (2017). "The State of Applied Econometrics: Causality and Policy Evaluation". *Journal of Economic Perspectives* 31.2, 3–32.

Backus, Matthew et al. (2020). "Sequential bargaining in the field: Evidence from millions of online bargaining interactions". *The Quarterly Journal of Economics* 135.3, 1319–1361.

Balakrishnan, Anantaram, Shankar Sundaresan, and Bo Zhang (2014). "Browse-and-switch: Retail-online competition under value uncertainty". *Production and Operations Management* 23.7, 1129–1145.

Bellotti, Victoria et al. (2017). "Why Users Disintermediate Peer-to-Peer Marketplaces". *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 4370–4382.

Binmore, Ken, Ariel Rubinstein, and Asher Wolinsky (1986). "The Nash bargaining solution in economic modelling". *The RAND Journal of Economics*, 176–188.

Chaves, Isaıas N (2018). "Pricing Responses to Platform Leakage: Optimal Platform Design When Matches Are Irrevocable". Unpublished manuscript.

Coase, Ronald Harry (1937). "The nature of the firm". *Economica* 4.16, 386–405.

Dworczak, Piotr, Scott Duke Kominers, and Mohammad Akbarpour (2021). "Redistribution through markets". *Econometrica* 89.4, 1665–1698.

eBay (2022a). *Buy or sell outside of eBay policy*. URL: https://www.ebay.com/help/policies/payment-policies/offering-buy-sell-outside-ebay-policy?id=4272.

— (2022b). *Member-to-member contact policy*. URL: https://www.ebay.com/help/policies/member-behavior-policies/membertomember-contact-policy?id=4262.

Edelman, Benjamin and Philip Hu (2016). "Disintermediation in two-sided marketplaces". *Harvard Business School Technical Note* 917.4.

Einav, Liran, Chiara Farronato, and Jonathan Levin (2016). "Peer-to-Peer Markets". *Annual Review of Economics* 8.1, 615–635.

Fong, Jessica (2020). "Search, selectivity, and market thickness in two-sided markets: Evidence from online dating". *SSRN 3458373*.

Galichon, Alfred and Bernard Salanié (2021). "Structural estimation of matching markets with transferable utility". *arXiv preprint arXiv:2109.07932*.

Gu, Grace and Feng Zhu (2021). "Trust and disintermediation: Evidence from an online freelance marketplace". *Management Science* 67.2, 794–807.

Hagiu, Andrei and Julian Wright (2021). *Platform leakage I. Platform Chronicles*. URL: https://platformchronicles.substack.com/p/platform-leakage.

— (2022). "Marketplace Leakage". Unpublished manuscript.

He, Eryn Juan et al. (2020). "Off-platform Threats in On-Demand Services". *SSRN 3550646*.
Jiang, Zhenling (2022). "An empirical bargaining model with left-digit bias: A study on auto loan monthly payments". *Management Science* 68.1, 442–465.
Jing, Bing (2018). "Showrooming and Webrooming: Information Externalities between Online and Offline Sellers". *Marketing Science* 37.3, 469–483.
Kuksov, Dmitri and Chenxi Liao (2018). "When showrooming increases retailer profit". *Journal of Marketing Research* 55.4, 459–473.
Ladd, Ted (2022). "The Achilles' heel of the platform business model: Disintermediation". *Business Horizons* 65.3, 277–289.
Manski, Charles F (2003). *Partial identification of probability distributions*. Vol. 5. Springer.
— (2007). "Partial identification of counterfactual choice probabilities". *International Economic Review* 48.4, 1393–1410.
McFadden, Daniel and Kenneth Train (2000). "Mixed MNL models for discrete response". *Journal of applied Econometrics* 15.5, 447–470.
Mehra, Amit, Subodha Kumar, and Jagmohan S Raju (2018). "Competitive strategies for brick-and-mortar stores to counter showrooming". *Management Science* 64.7, 3076–3090.
Nash, John F (1950). "The bargaining problem". *Econometrica: Journal of the econometric society*, 155–162.
Peitz, Martin and Anton Sobolev (2022). "Inflated recommendations". CEPR Discussion Paper No. DP17260.
Rochet, Jean-Charles and Jean Tirole (2006). "Two-sided markets: a progress report". *The RAND journal of economics* 37.3, 645–667.
Rubinstein, Ariel (1982). "Perfect equilibrium in a bargaining model". *Econometrica: Journal of the Econometric Society*, 97–109.
Shapiro, Susan P (1987). "The social control of impersonal trust". *American journal of Sociology* 93.3, 623–658.
Shen, Jill (2019). *Didi faces new privacy pressures with mandatory audio recording*. URL: `https://technode.com/2019/12/24/didi-carpool-audio-recording`.
Sieg, Holger (2000). "Estimating a bargaining model with asymmetric information: Evidence from medical malpractice disputes". *Journal of Political Economy* 108.5, 1006–1021.
Spulber, Daniel F (2019). "The economics of markets and platforms". *Journal of Economics & Management Strategy* 28.1, 159–172.
Uber (2022). *Street Hails and Off-platform Pickups*. URL: `https://www.uber.com/legal/en/document/?name=general-community-guidelines&country=united-states`.
Wang, Chengsi and Julian Wright (2020). "Search platforms: showrooming and price parity clauses". *RAND Journal of Economics* 51.1, 32–58.
Weyl, E. Glen (2010). "A Price Theory of Multi-sided Platforms". *American Economic Review* 100.4, 1642–72.
Williamson, Oliver E (1987). "Transaction cost economics: The comparative contracting perspective". *Journal of economic behavior & organization* 8.4, 617–625.
Williamson, William and James Scrofani (n.d.). "Trends in detection and characterization of propaganda bots". *Proceedings of the 52nd International Conference on System Sciences*.
Wu, Dazhong et al. (2004). "Implications of reduced search cost and free riding in e-commerce". *Marketing Science* 23.2, 255–262.

Xie, Yingkang, Jingyi Wang, and Huaiyu Zhu (2022). “Monitoring Disintermediation: Actions Matter More Than Words”. Working Paper, Kellogg School of Management.

Zhang, Xu, Junhong Chu, and Puneet Manchanda (2021). “Online Healthcare Platform Evolution: The Interplay of Bargaining and Network Effects”. *SSRN 3965007*.

Zhou, Qiang (Kris) et al. (2022). “Platform Exploitation: When Service Agents Defect with Customers from Online Service Platforms”. *Journal of Marketing* 86.2, 105–125.

Zhu, Feng, Weiru Chen, and Shirley Sun (2018). “ZBJ: Building a Global Outsourcing Platform for Knowledge Workers”. *Harvard Business School Case* 618-044.

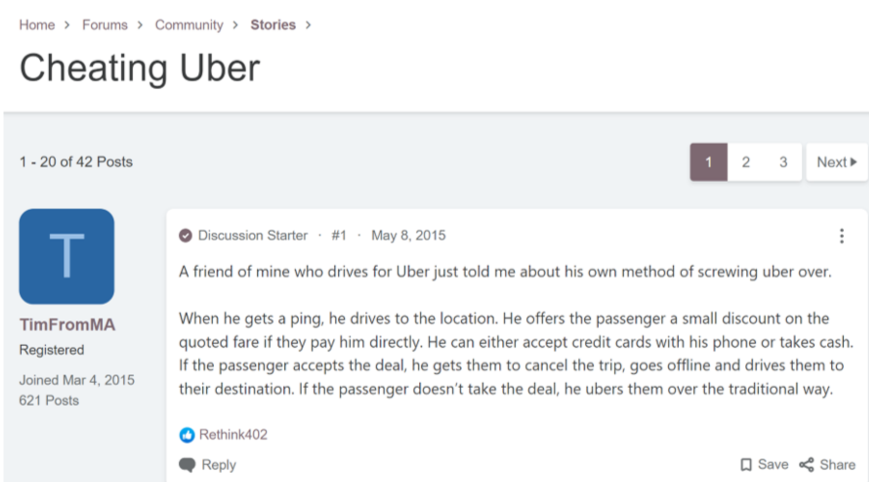
Appendix

A Industry Overview

Platform leakage is common in the gig economy, which involves connecting independent contractors with customers for flexible and temporary jobs through an online platform. The requirement of individual interaction is one reason why many emerging platforms built for independent contractors might have been hard hit by disintermediation[42] [43] [44]. Services are fulfilled directly by the workers, as is the communication with customers. Leakage can happen as individuals share information, gain trust, and make direct payments. Digital platforms with peer-to-peer transactions (Einav et al., 2016) are vulnerable to offline collusion.

In on-demand service platforms, it is not uncommon for rideshare and delivery drivers to deal directly with their customers. A collection of survey and interview studies on U.S. residents (Bellotti et al., 2017) documents the prevalence of disintermediation, including multiple cases of cash deals where Uber and Lyft drivers asked the customer to cancel the trip on the app and continue the trip outside the platform. On a site for rideshare drivers, people discussed providing customers with a discount for offline transactions (see Figure 11).

Figure 11: The Discussion of Disintermediation on UberPeople.net



Note: Forum users shared how drivers can disintermediate the platform by offering a discount to the customer. (`https://www.uberpeople.net/threads/cheating-uber.19545/`)

However, the extent of leakage varies widely across marketplaces. A higher price increases the absolute size of the commission (even if it is a low percentage), which raises the savings if the buyer and seller bypass the platform (Edelman and Hu, 2016). The monetary incentive of avoiding commission is larger in cargo delivery services (e.g, Lalamove, Manbang, Convoy, Dolly) than in other types of services such as ride-sharing (e.g., Uber, Lyft, Didi) and food

[42] Madden, Sam (2015).Why Homejoy Failed And The Future Of The On-Demand Economy URL:https://techcrunch.com/2015/07/31/why-homejoy-failed-and-the-future-of-the-on-demand-economy

[43] Said, Carolyn (2015). Could Client Poaching Undercut On-Demand Companies? URL:www.sfchronicle.com/business/article/Could-client-poaching-undercut-on-demand-6222919.php

[44] Sarva, Amol and Jeff Wald (2015). The Missing Leak In Marketplaces URL:www.techcrunch.com/2015/12/16/the-missing-leak-in-marketplaces

delivery (e.g., Doordash, Instacart, Postmates, Meituan), because the revenue per job are higher for a longer trip and a larger size of delivery in general. For example, a full truckload of items for a 15-mile delivery would be more expensive than the typical food delivery within a five-mile radius. The same 15% commission rate would thus have different implications.

Therefore, we choose to investigate the leakage problem in on-demand cargo delivery services, which are possibly vulnerable to disintermediation due to high commission savings.

B Detection Algorithms

The detection algorithms provide evidence of offline transactions. We develop and compare two algorithms that use geolocation data and conversation data. A team of human analysts labeled a random sample of more than 5000 cancellations in parallel to the algorithmic classification. They verified offline transactions using GPS tracking information through the driver's app, and further corroborated using information such as text messages and phone call logs. Using the human labels as ground truth, we validate that the GPS detection algorithm can achieve both high recall and precision.

B.1 GPS Detection (Direct Measure)

We checked whether the driver passed the origin and destination of a previously canceled delivery. The assumption is that, if a job was taken offline, the driver would still visit the origin and destination within the previously agreed upon time window for the requested trip.

The key data used in this algorithm are the uploaded GPS track points when a driver uses the mobile app or has the app service running in the background on their cell phones. Presumably, the app is active for drivers who are looking for jobs either when they are idle or occupied by the last job. The platform broadcasts a new job to drivers who are nearby or would be arriving at the address of the request; hence, drivers have an incentive to keep the platform informed about their geolocations. The app obtained drivers' consent for using the geolocation data for business operations. There were no penalties for disintermediation because the platform didn't have ways to verify offline transactions, and thus drivers don't have incentives to hide their location other than personal privacy concerns.

To reduce the computation cost in mining the large-scale GPS streams, we focus on drivers' footprints within a reasonable time window around the time of service specified in a canceled request. Since 90% of shipping requests are on-demand, it is not likely that the actual shipment would take place a day later. The algorithm calculates the distances between drivers' GPS track points and the shipment addresses to check whether the driver is present within a small radius of both the loading and unloading locations specified by the customer. The time window and the small radius[45] is tuned through offline iterative process using past labeled cancellations before the algorithm is officially deployed in production.

Admittedly, misclassification can happen when geolocation data is incomplete or incorrect. For example, drivers could turn off their GPS or completely shut down the phone. Poor cell phone signals might also affect the trackpoint uploads of some drivers. However,

[45] To maintain the proprietary nature of the algorithm, we can't disclose the specific values of time and distance criteria. The information could be used by drivers or customers to game against the system.

given the scalability and explainablity of the GPS detection, we accept its flaw as long as the algorithm delivers performance that is on par with human analysts.

B.2 Keyword Detection (Indirect Measure)

Monitoring conversations is a common and established practice (Airbnb, 2022b; eBay, 2022b; Gu and Zhu, 2021) to detect leakage in both industry and academia. We use keywords to check if users exchange contact information or propose offline coordination. Mentioning the names of alternative payment systems and other communication tools indicate the intention to disintermediate. When drivers or customers ask for private deals or describe the intention to avoid fees, leakage might happen. Besides using the relevant keywords, we also look for the sequence of numbers (e.g., three or more consecutive digits), which could be the substrings of phone numbers for making a call or account search in other systems. In summary, we perform simple "OR" statements of sensitive keywords and linguistic patterns using the regular expression to classify cancellations.

B.3 Human Evaluation

We implement the GPS detection in production. To evaluate the real-world performance of the algorithm, a team of human analysts independently label a new set of randomly selected 5557 cancellations. Human analysts confirmed 1009 disintermediated transactions out of these cancellations. The empirical evaluation follows the same process described in (Xie et al., 2022), in which human analysts review the activities in a cancelled trip, listen to the call conversations between driver and customer, and check the GPS footprints of the driver.

Table 5: The Performance of Detection Solutions

Data	Algorithm	Precision	Recall	F1-Score
Conversation	Keyword Detection	38.00%	58.67%	0.461
Geolocation	GPS Detection	90.38%	94.05%	0.922
Conversation + Geolocation	Combined Matching	49.05%	100.00%	0.658

Note: precision is the percentage of algorithmic classified leakage that are confirmed by human; recall (hit rate) is the percentage of all human confirmed leakage that are retrieved by the algorithm; F1 score is the harmonic precision-recall mean.

Using the human labels as ground truth, we report the precision, recall, and F1-score of the algorithm in Table 5. The GPS detection can recover 94.05% of all offline transactions labeled by human analysts and achieve a precision of 90.38% for correctly identifying leakage. We validate that the algorithm using only geolocation data has both high precision and recall, albeit with no access to soft information in conversations and cancellation reasons. In fact, we find that combining the GPS detection with keyword matching in conversation data would drop the precision from 90.38% to 49.05% with minor improvement in the recall.

Text mining by itself can miss more than 40% of the offline transactions. As a result, firms that only rely on conversation data may significantly underestimate the scale of leakage. Xie et al. (2022) also explore the Bidirectional Encoder Representations from Transformers (BERT), which is the state-of-the-art model in natural language processing. However, results show that the BERT solution is dominated by the GPS detection with minimal improvement in the precision that comes with an expensive tradeoff of dropping recall. Text mining is not reliable in leakage detection, perhaps because people can encrypt their contact exchanges (e.g., add random characters or sounds in between phone numbers) or have their conversations outside the platform (e.g., chat in person or connect on social networks).

C Synthetic Control Methods (SCM)

C.1 The SCM Estimates of Beijing

We estimate the city-specific treatment effects on Non-VIPs using synthetic controls for 33 cities in Section C.2. The synthetic control method (Abadie et al., 2010) is "arguably the most important innovation in the evaluation literature in the last fifteen years" (Athey and Imbens, 2017). The data-driven approach constructs a measure of the counterfactual leakage rate for each treated city that would have occurred had the city not charged a commission fee. In this section, I will use Beijing as an example to illustrate how I implement the synthetic controls estimation. In Section D.1, we will compare this city-level estimate with the driver-level estimates that tracks the same set of drivers overtime.

First, I define the set of candidate controls as the set of all cities without any treatment changes within the 28 days before and 28 days after the launch date of Beijing. Then, using the data in the pre-period, I estimate the parameters of synthetic controls using a constrained linear regression. Let y_{ct} represent the daily leakage rate of the treated city c (e.g., Beijing) on day t, and let T_0 denote the set of days in the pre-period. Let $K \in K$ index the set of candidate control cities. Finally, I estimate a vector of weights $w = \{w_k\}_{k \in K}$ by minimizing the following objective function:

$$\begin{aligned} \underset{w}{min} \sum_{t \in T_0} \left(y_{ct} - \sum_{k \in K} w_k y_{kt} \right)^2 \\ s.t. \sum_{k \in K} w_k = 1, w_k \geq 0 \ \forall k \in K \end{aligned} \tag{14}$$

Weights $\hat{w}$ are chosen so that the synthetic control group's pre-period leakage rate closely matches the average of the treated groups. The intuition is that some cities (e.g., neighboring cities) are more similar to the treated city than others in the entire country, and those cities should contribute more to the estimate of the counterfactual leakage rate in the treated city.

Using the weights in $\hat{w}$, I can construct the measure of the fitted leakage rate in the pre-period T_0 and the measure of the counterfactual leakage rage had Beijing not charged a commission fee in post-period T_1. The dash line in Figure 12 demonstrates the synthetic controls counterfactual $\sum_{k \in K} \hat{w}_k y_{kt}$ for each t in both T_0 and T_1. The gaps between the counterfactual and actual leakage rate inform us about the treated effect each day. The city-specific treatment effect for the post-period with $|T_1| = 28$ days is thus:

$$\frac{1}{|T_1|} \sum_{t \in T_1} \left(y_{ct} - \sum_{k \in K} \hat{w}_k y_{kt} \right)^2 \tag{15}$$

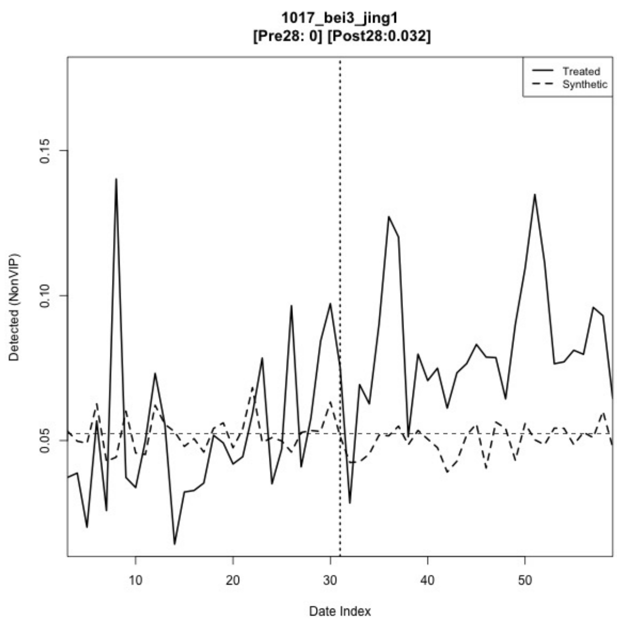


Figure 12: Detected Rate

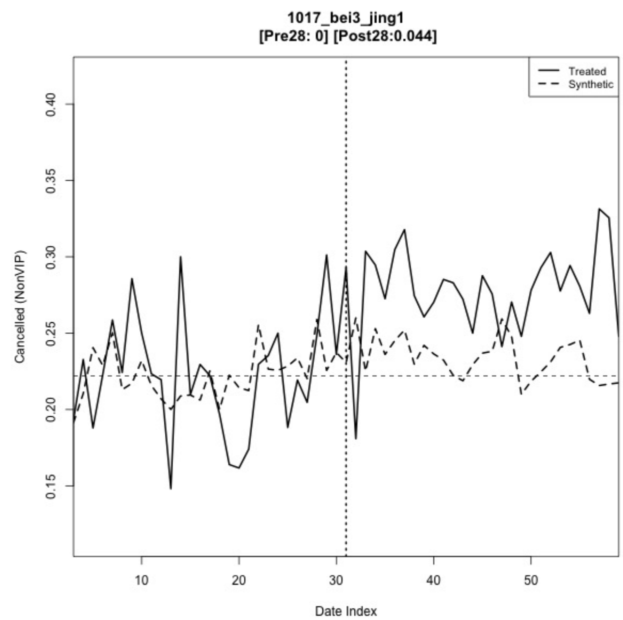


Figure 13: Cancellation Rate

The basic idea of the synthetic control method is to utilize pre-period data to construct weighted averages of the non-treated units that fit the treated unit well, and then use those weights to construct the counterfactual for each treated unit in the post-period. Figure 12 shows that we have a good fitting with the average difference between the fitted and actual leakage rate at zero for $|T_0| = 28$ days before the intervention.

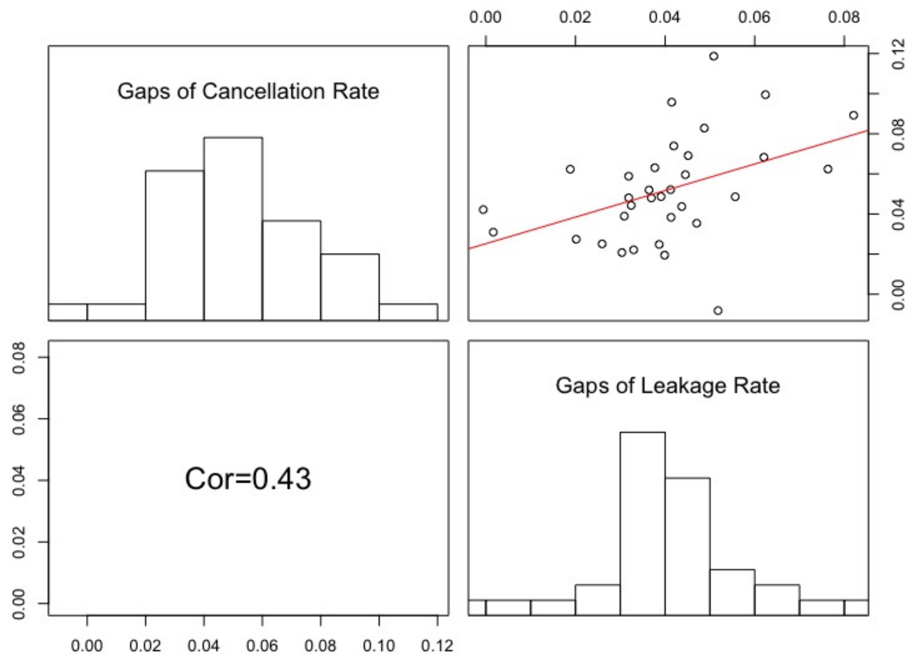
The estimated Beijing-specific treatment effect on leakage rate is 3.2% in the 28 days (see Figure 12) after the intervention. We can conduct the same exercise for cancellation rate and find that 4.4% additional transactions are canceled in the 28 days (see Figure 13)after charging the commission fee.

C.2 The Distribution of SCM Estimates of All Treated Cities

We are interested in an overview of the treatment effects across different cities. The information helps inform the sampling strategy for the structural model in Section 4. If we see obvious heterogeneity across cities, we should randomly draw drivers from all over the country to obtain a representative sample to generate insights for the platform. I choose synthetic control methods over difference-in-difference regressions because I can run the estimations in a loop without manually finding adjacent cities for the control group.

We estimate the city-specific treatment effects by creating synthetic controls using Equation (14) for all 33 cities that are treated in the 137 days. Figure 14 demonstrates the city-specific treatment effects on leakage rate and cancellation rate for Non-VIPs. The treatment effects are the gaps between the counterfactual and actual leakage rate. They are calculated based on Equation (15) in the post-period. The top right panel shows that treatment effects on the two metrics are positively correlated and mostly non-negative.

Figure 14: The City-Specific Treatment Effects for Non-VIPs from Synthetic Control Method



We can conduct similar exercise for Super-VIPs. Table 6 documents the summary statistics of city-specific treatment effects for both the Non-VIPs and Super-VIPs in the 33 cities. The cancellation rate is a nosier metric than the leakage rate according to the standard deviation. This observation can be confirmed by looking into the histograms (Figure 14) of treatment effects on leakage rate are more concentrated than the treatment effects on cancellation rate, or by looking into the wider confidence intervals in Figure 16 than the ones in Figure 15.

Table 6: Fitted and Predicted Gaps of Synthetic Controls

	Metric	Fitted Gaps Pre28 Mean	Predicted Gaps Post28 Mean (Std Dev)
Non-VIP	% Detected	0%	+4.00% (1.71%)
	% Cancelled	0%	+5.17% (2.65%)
Super-VIP	% Detected	0%	-0.19% (0.34%)
	% Cancelled	0%	-1.19% (1.24%)

Moving on from demonstrating the heterogeneity across cities, Figure 15 averages across the 33 cities and shows the mean detected ratio per day by centering their time series at the launch date. The increase in average leakage rate for Non-VIP drivers after the commission launch is prominent. In contrast, we do not see any visible changes in the average leakage rate for VIP drivers who were not charged the 15% commission rates. Alternatively, we can check how cancellations change after the launch of the commission. Figure 16 shows the average cancellation rate per day across the 33 cities by centering their time series at the launch date. The mean pre- and post-intervention cancellation rates are prominent again for Non-VIP drivers but not for VIP drivers who had fee waiver.

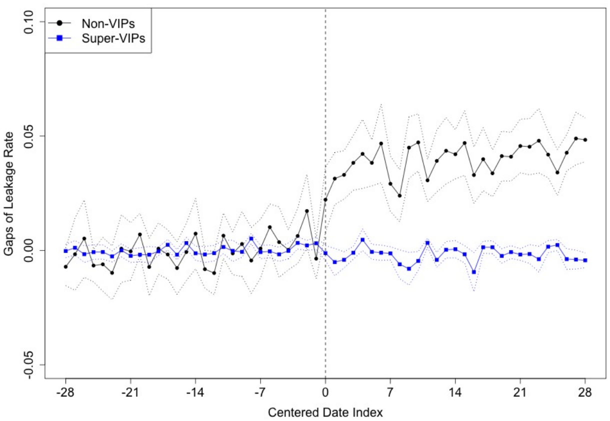


Figure 15: Effects on Detected Rate

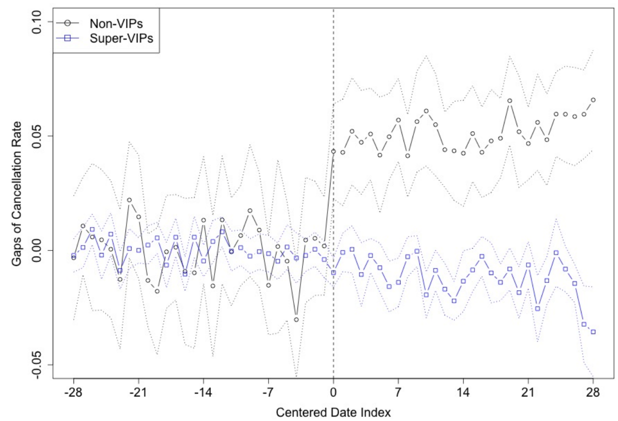


Figure 16: Effects on Cancellation Rate

The histograms of estimates in Figure 14 from synthetic control methods demonstrate the variation of commission effects. While the variation of the effects on cancellation rates is larger than the ones on leakage rate, both city-level effects are positively correlated and scattered around the diagonal line. Using additional geolocation data seems to reduce the noise for verifying disintermediated transaction.

Another source of the variation in city-level effects might come from the heterogeneity across cities - different markets feature a different mix of driver composition. With the suggestive evidence in mind, we decide to randomly draw samples of drivers from the entire country to leverage the potential heterogeneity in fee sensitivity across geographical regions because it provides identification for the primitives in the structural model in Section 4.

D Difference-in-Difference (DiD)

In Section C.1, I use synthetic control methods to estimate the city-specific treatment effect for Beijing with the aggregate transaction data. Although the city-level data covers the universe of transactions, synthetic control methods cannot solve the problem of changes in driver composition. As a complementary approach, I will conduct the driver-level difference-in-difference (DiD) regression that tracks the same subset of drivers in this section.

D.1 Driver-level DiD Estimates of Beijing

We limit the analysis to Beijing, Tianjin and Langfang that are adjacent cities. By tracking the same drivers overtime, we can isolate the within-individual changes before and after from the changes in driver composition. Secondly, drivers in Langfang can serve as DiD control group, because they are not affected by the commission until four weeks after Beijing. Lastly, the unanticipated[46] introduction of commission is exogenous because different VIP expiration dates shift the incentive shock across drivers in the same city.

With the identification strategies listed above, we pull all available sample of drivers in the Jing-jin-ji Metropolitan Region and track their daily business metrics (e.g., leakage rate and cancellation rate) for 28 days before and 28 days after the launch of commission (i.e. $PolicyV2$) in Beijing and Tianjin. We limit the analysis to drivers that are both active before and after the commission shock to leverage the with-in individual variation. The resulting samples have 12071 Beijing drivers, 4186 Tianjin drivers, and 2068 Langfang drivers. In total, they contributed 540,415 job assignments on the platform during these 57 days. We can run a reduced-form analysis with the following driver-level regression:

$$Y_{ict} = \alpha_i + \beta \cdot X_{ict} + \beta_v \cdot NonVIP_{ict} + \beta_p \cdot PolicyV2_{ict} + \delta \cdot PolicyV2_{ict} \times NonVIP_{ict} + \epsilon_{ict} \quad (16)$$

where Y_{ict} denotes the leakage rate or cancellation rate for driver i in city c on day t. α_is are the driver fixed effects that represent the baseline leakage or cancellation rate for driver i, X_{ict} are the control variables that contain week fixed effects, day-of-week dummies, and city fixed effects. The $PolicyV2_{ict}$ is a dummy equal to one if the new policy of commission is launched at city c (e.g., only Beijing drivers during the post-period would have $PolicyV2_{ict} = 1$). $NonVIP_{it}$ is a dummy variable that represents whether the driver is Non-VIP or Super-VIP.

Table 7: Driver-level DiD Regressions (28 Days Before and After)

	Leakage Rate		Cancellation Rate	
	(1)	(2)	(3)	(4)
PolicyV2	0.0004	0.002	-0.022***	-0.010***
	(0.002)	(0.001)	(0.006)	(0.004)
NonVIP	0.015***	0.019***	0.177***	0.069***
	(0.001)	(0.002)	(0.008)	(0.006)
PolicyV2 × NonVIP	0.027***	0.023***	0.013	0.033***
	(0.004)	(0.003)	(0.010)	(0.007)
Driver F.E.		✓		✓
Week-of-year F.E.	✓	✓	✓	✓
Day-of-week F.E.	✓	✓	✓	✓
City (Beijing, Tianjin, Langfang)	✓	✓	✓	✓
Observations	540,415	540,415	540,415	540,415
R^2	0.00663	0.08033	0.01861	0.14309
Note:			*p<0.1; **p<0.05; ***p<0.01	

Table 7 shows that the treatment effects of commission on Non-VIP drivers are 2.3% for leakage and 3.3% for cancellation. These results are estimated from the data with drivers who

[46]The announcement of commission fee on Non-VIPs is made a week before the official launch. Drivers who paid for the membership a month ago will have their VIP status expiring randomly in the next month.

are both active 28 days before and after the commission shock. They are smaller than the 3.2% and 4.4% estimated using the synthetic control methods with all available transactions.

The two approaches, the difference-in-difference (DiD) regression at the driver level and the synthetic control methods at the city level, have different sources of bias. The treatment effects of 2.3% for leakage and 3.3% for cancellation, which are estimated by the driver-level DiD regression, can only apply to drivers who are active 28 days before and 28 days after the commission shock. These active drivers might be less sensitive to commission fees. In contrast, the city-level synthetic control methods make use of all available transactions. However, it can't solve the driver composition problem: drivers with the high costs to disintermediate may choose to subscribe to the VIP status, and drivers that are inherently more likely to disintermediate stay as the Non-VIPs. Therefore, both approaches come with different pros and cons.

D.2 Discussion: Suggestive vs. Causal Evidence

Both approaches in Section C.1 (synthetic control unit at the city level) and Section D.1 (individuals in an adjacent city at the driver level) ultimately assume parallel trends for the constructed control and treated: the leakage rate in the treated group (cities or drivers) would have been the same as in the control group but for the commission launch. There are pros and cons of the two methods: (1) the city-level synthetic control methods use the universe of transactions and borrow variation from other cities with potential bias caused by the changes in driver composition, and (2) the driver-level regressions can track the same drivers overtime with the sampling bias introduced by only studying the active drivers.

While researchers want to reduce bias by developing new identification strategies or only make claims on the average treatment effects of a subset of drivers, practitioners are more interested in a quick overview of city-specific treatment effects across different cities. More importantly, I am interested in examining whether there is heterogeneity across cities. Therefore, I decided to put the city-level synthetic control methods in the main text of the paper and have the driver-level DiD regression in this appendix.

E Intuition from Partial Identification

We can use partial identification (Manski, 2003; Manski, 2007) to compute bounds that summarize what the data say about the individual parameters. The following exercise helps us understand the identification strategy for the structural model.

Decision Rule

$$L_{i(t)} = 1 \quad \text{iff} \quad \gamma_i p_t - \lambda_{i(t)} p_t > h_{i(t)} \tag{17}$$

Equation (17) describes that driver i wants to take transaction t of the platform if and only if the online commission exceeds the offline friction. This is the simplified version of Equation (3) in Section 4.2.1. To focus on the intuition of identification, we make the following assumptions to simplify the statements of inequality: (1) assume customers are i.i.d on transactions; (2) assume drivers make a take-it-or-leave-it offer where the discount only varies by transactions; (3) assume $\beta_i = 1$ (i.e., normalizing $h_{i(t)}$ by β_i to have the same unit as commission fees and quoted price). The above assumptions don't reflect the real world but create examples for us to understand the basic intuition of the identification.

Data Points and Identification

In the data, we observe the following three data points:

1. Non-leakage on transaction t with $\gamma_{it} = 0$

$$L_{i(t)} = 0 \quad \text{iff} \quad 0 - \lambda_{i(t)} \leq h_{i(t)} \tag{18}$$

2. Non-leakage on transaction $t^{'}$ with $\gamma_{it'} = 0.15$

$$L_{i(t')} = 0 \quad \text{iff} \quad 0.15p_{t'} - \lambda_{i(t')} > h_{i(t')} \tag{19}$$

3. Leakage on transaction $t^{''}$ with $\gamma_{it''} = 0.15$

$$L_{i(t'')} = 1 \quad \text{iff} \quad 0.15p_{t''} - \lambda_{i(t'')} > h_{i(t'')} \tag{20}$$

First, let's focus on backing out the hassle for driver i at the time of transacting in transaction t. To get the basic intuition, we make one last assumption: $\lambda_{i(t)} \approx 0$ for very small p_t, where drivers don't bother to give any discount in very low-value jobs.

The combinations of the above data points identify h_i and λ_i:

- Within-unit price variation: same driver seeing different prices after shock
  - Eq(19) and Eq(20): $h_i \in [0.15p_{t'}, 0.15p_{t''}]$
- Within-unit shock variation: same driver seeing same price before and after
  - Eq(18) and Eq(19): $\lambda_{i)} \in [\frac{0.15p_t - 2h_i}{2p_t}, \infty]$
  - Eq(18) and Eq(20): $\lambda_{i)} \in [-\frac{h_i}{p_t}, 0.15 - \frac{h_i}{p_t}]$
- Within-unit price and shock variation: different prices before and after the shock
  - Eq(18) and Eq(19): $\lambda_{i)} \in [\frac{0.15p_{t'} - 2h_i}{p_t + p_{t'}}, \infty]$
  - Eq(18) and Eq(20): $\lambda_{i)} \in [0, \min\{0.15 - \frac{h_i}{p_{t''}}, \frac{0.15p_{t''} - 2h_i}{p_{t''} - p_t}\}]$
  - Eq(19) and Eq(20): $\lambda_{i)} \in [0.15 - \frac{h_i}{p_t^{'}}, 0.15 - \frac{h_i}{p_t^{''}}]$

To summarize, the driver-side hassle is partially identified by within-driver price variation after shock with assumptions. The individual preferred discount is partially identified by the combination of within-unit commission shock and price variation. We can apply similar intuition to our identification strategies in the structural model.

F Additional Tables of Structural MLE Estimates

Table 8: Homogeneous Price Sensitivity within Active Customers

	Dependent variable: Disintermediation			
	(1)	(2)	(3)	(4)
commission_fee ($\gamma_{i(t)}p_t$)	0.345***	0.279***	0.357***	0.320**
	(0.112)	(0.090)	(0.126)	(0.125)
subsidy_to_customer ($s_{j(t)}$)	-0.119	-0.082	0.006	0.026
	(0.297)	(0.298)	(0.481)	(0.481)
transaction_price (p_t)	-0.046***	-0.043***	-0.058***	-0.057***
	(0.011)	(0.011)	(0.018)	(0.018)
is_cash	1.50***	1.46***	2.08***	2.06***
	(0.325)	(0.325)	(0.651)	(0.651)
is_furniture	-7.60***	-7.62***	-263.6	-316.6
	(2.10)	(2.10)	(178.1)	(231.9)
is_scheduled	0.119	0.127	0.691	0.688
	(0.380)	(0.375)	(0.638)	(0.639)
is_intercity	0.409	0.368	0.800	0.762
	(0.320)	(0.320)	(0.581)	(0.590)
is_bus_ep	0.382	0.268	-0.360	-0.411
	(1.56)	(1.52)	(2.01)	(2.04)
passenger1	-0.246	-0.259	-0.612	-0.605
	(0.474)	(0.470)	(0.647)	(0.646)
passenger2	-0.530	-0.499	-2.01*	-2.05*
	(0.629)	(0.627)	(1.14)	(1.17)
vehicleTruck_M	1.05***	0.858**	6.20**	6.02**
	(0.375)	(0.375)	(2.51)	(2.34)
vehicleTruck_S	0.496**	0.355	0.462	0.346
	(0.245)	(0.248)	(1.10)	(1.10)
vehicleVan_M	0.263	0.220	0.530	0.427
	(0.235)	(0.238)	(0.767)	(0.758)
vipDriver		-1.03***		-1.41*
		(0.304)		(0.775)
Fixed-effects: driver_id			✓	✓
Fixed-effects: user_id	✓	✓	✓	✓
Average Discount ($\bar{\lambda}$)	0.205	0.216	0.160	0.165
Observations	3,336	3,336	3,244	3,244
- Unique Users	989	989	989	989
- Unique Drivers	912	912	842	842
Pseudo R^2	0.19228	0.19778	0.58839	0.59040
BIC	11,510.9	11,496.0	16,597.7	16,597.5

Table 9: Heterogeneous Price Sensitivity and Differential Discounts

	Dependent variable: Disintermediation			
	(1)	(2)	(3)	(4)
commission_fee ($\gamma_{i(t)}p_t$)	0.636***	0.509***	0.595***	0.514***
	(0.043)	(0.046)	(0.062)	(0.064)
subsidy_to_customer ($s_{j(t)}$)	-0.016	-0.221*	-0.009	-0.204
	(0.077)	(0.130)	(0.084)	(0.143)
transaction_price (p_t)	-0.020***	-0.010**	-0.015***	-0.006
	(0.004)	(0.005)	(0.005)	(0.006)
is_cash	1.14***	1.15***	1.16***	1.17***
	(0.042)	(0.042)	(0.062)	(0.061)
is_furniture	-3.06***	-3.06***	-3.12***	-3.12***
	(0.228)	(0.232)	(0.222)	(0.228)
is_scheduled	0.505***	0.487***	0.442***	0.433***
	(0.061)	(0.061)	(0.066)	(0.065)
is_intercity	-0.274***	-0.245***	-0.180**	-0.165**
	(0.069)	(0.068)	(0.084)	(0.081)
is_bus_ep	-0.296**	-0.275*	-0.348**	-0.355**
	(0.147)	(0.149)	(0.159)	(0.159)
passenger1	0.094	0.101	0.080	0.084
	(0.068)	(0.068)	(0.072)	(0.072)
passenger2	0.304***	0.316***	0.316***	0.323***
	(0.087)	(0.087)	(0.098)	(0.097)
vehicleTruck_M	0.288***	0.243***	0.174	0.178
	(0.077)	(0.076)	(0.205)	(0.199)
vehicleTruck_S	0.225***	0.195***	0.090	0.088
	(0.057)	(0.056)	(0.126)	(0.124)
vehicleVan_M	0.121*	0.122*	0.065	0.066
	(0.066)	(0.066)	(0.099)	(0.095)
commission_fee × is_cash	-0.220***	-0.213***	-0.196***	-0.201***
	(0.030)	(0.029)	(0.052)	(0.044)
commission_fee × is_furniture	-18.2	-18.4	-18.8***	-18.8***
	(119.0)	(119.5)	(0.441)	(0.449)
commission_fee × is_scheduled	-0.096***	-0.077**	-0.076	-0.063
	(0.036)	(0.036)	(0.050)	(0.047)
commission_fee × is_intercity	-0.280***	-0.230***	-0.206***	-0.187***
	(0.031)	(0.031)	(0.050)	(0.046)
commission_fee × passenger1	-0.041	-0.035	-0.051	-0.052
	(0.053)	(0.054)	(0.067)	(0.065)
commission_fee × passenger2	-0.018	-0.005	-0.033	-0.021
	(0.048)	(0.046)	(0.066)	(0.059)
commission_fee × vehicleTruck_M	-0.110**	-0.067	-0.145**	-0.116**
	(0.044)	(0.044)	(0.062)	(0.059)

commission_fee × vehicleTruck_S	0.003	0.032	-0.067	-0.045
	(0.041)	(0.042)	(0.054)	(0.052)
commission_fee × vehicleVan_M	-0.075	-0.040	-0.152	-0.101
	(0.063)	(0.057)	(0.161)	(0.094)
subsidy_to_customer × is_furniture	-0.047	-0.057	-0.097	-0.103
	(0.497)	(0.497)	(0.510)	(0.508)
subsidy_to_customer × is_scheduled	0.094	0.105	0.107	0.119
	(0.113)	(0.113)	(0.124)	(0.123)
subsidy_to_customer × is_intercity	0.080	0.088	0.037	0.039
	(0.130)	(0.129)	(0.137)	(0.134)
subsidy_to_customer × is_bus_ep	0.022	0.020	-0.009	-0.013
	(0.114)	(0.114)	(0.119)	(0.118)
subsidy_to_customer × passenger1	-0.017	-0.020	-0.002	-0.001
	(0.168)	(0.168)	(0.188)	(0.186)
subsidy_to_customer × passenger2	-0.020	-0.020	-0.066	-0.077
	(0.191)	(0.191)	(0.194)	(0.194)
subsidy_to_customer × vehicleTruck_M	0.027	0.059	-0.003	0.018
	(0.130)	(0.131)	(0.148)	(0.150)
subsidy_to_customer × vehicleTruck_S	-0.015	0.009	0.008	0.042
	(0.114)	(0.115)	(0.130)	(0.130)
subsidy_to_customer × vehicleVan_M	-0.345**	-0.342**	-0.334**	-0.327**
	(0.162)	(0.161)	(0.142)	(0.141)
transaction_price × is_cash	-0.008***	-0.009***	-0.009***	-0.010***
	(0.003)	(0.003)	(0.003)	(0.003)
transaction_price × is_furniture	0.008	0.008	0.009	0.009
	(0.012)	(0.012)	(0.009)	(0.010)
transaction_price × is_scheduled	-0.004	-0.004	-0.004	-0.004
	(0.003)	(0.003)	(0.003)	(0.003)
transaction_price × is_intercity	0.007**	0.004	0.004	0.003
	(0.004)	(0.003)	(0.004)	(0.004)
transaction_price × is_bus_ep	0.008*	0.005	0.012**	0.012**
	(0.005)	(0.005)	(0.005)	(0.005)
transaction_price × passenger1	0.010***	0.010***	0.009**	0.008**
	(0.004)	(0.004)	(0.004)	(0.004)
transaction_price × passenger2	0.003	0.002	0.001	0.0004
	(0.005)	(0.005)	(0.006)	(0.005)
transaction_price × vehicleTruck_M	0.013***	0.011***	0.010**	0.009*
	(0.004)	(0.004)	(0.005)	(0.005)
transaction_price × vehicleTruck_S	0.003	0.002	0.001	0.0005
	(0.004)	(0.004)	(0.006)	(0.005)
transaction_price × vehicleVan_M	0.004	0.003	0.002	0.0009
	(0.005)	(0.005)	(0.006)	(0.006)
vipDriver		-0.319***		-0.333***
		(0.051)		(0.089)
commission_fee × vipDriver		-0.029		-0.013

		(0.034)		(0.051)
subsidy_to_customer × vipDriver		0.236**		0.224*
		(0.120)		(0.130)
transaction_price × vipDriver		-0.007***		-0.010***
		(0.003)		(0.003)
(Intercept)	-4.09***	-3.82***		
	(0.038)	(0.057)		
Fixed-effects: driver_id			✓	✓
Baseline Discount ($\bar{\lambda}$)	0.032	0.035	0.026	0.018
Observations	269,921	269,911	248,556	248,556
- Unique Drivers	1971	1962	1280	1280
- Unique Users	239,057	239,048	220,920	220,920
Pseudo R^2	0.05513	0.05718	0.11237	0.11349
BIC	53,226.6	53,162.3	65,057.0	65,045.1

Note: *p<0.1; **p<0.05; ***p<0.01

G Microfounded Bargaining

G.1 Nash Bargaining

Eq (8) assumes that driver i and customer j can reach an equilibrium outcome $\lambda^*_{ij(t)}$ when the joint utility gain from leakage is non-negative. The $\lambda^*_{ij(t)}$ enables agreement as long as both of their individual latent utility gains are non-negative.

This section uses the Nash bargaining model(Nash, 1950) to characterize the static equilibrium. The Nash bargaining solution is a convenient way to characterize the outcome when we do not observe the bargaining process(Jiang, 2022). Given the payment division rule $\lambda_{ij(t)} \in [0,1]$, the latent utility gains from leakage for driver i and customer j are:

$$\begin{aligned}\Delta\pi_{ij(t)} &= \Pi^1_{i(t)} - \Pi^0_{i(t)} \\ &= \beta_i \cdot (\gamma_{i(t)} - \lambda_{ij(t)})p_t - h_{i(t)}\end{aligned} \tag{21}$$

and

$$\begin{aligned}\Delta u_{ji(t)} &= U^1_{j(t)} - U^0_{j(t)} \\ &= \beta_j \cdot (\lambda_{ij(t)}p_t - s_{j(t)}) - h_{j(t)}\end{aligned} \tag{22}$$

The Nash bargaining solution (Sieg, 2000; Zhang et al., 2021) is given by maximizing the generalized Nash product specified by the bargaining weight η:

$$\begin{aligned}&\max_{\lambda_{ij(t)}} (\Delta\pi_{ij(t)})^{\eta} \cdot (\Delta u_{ji(t)})^{1-\eta} \\ &\text{s.t.} \quad \Delta\pi_{ij(t)} \geq 0 \\ &\qquad\;\; \Delta u_{ji(t)} \geq 0 \\ &\qquad\;\; 0 \leq \eta \leq 1\end{aligned} \tag{23}$$

where $\eta \in [0, 1]$ represent the relative bargaining power of driver.

We assume homogeneity of bargaining power in this analysis to start with a parsimonious model, which is micro-founded with fewer primitives. Future analyses can account for the heterogeneity in η across different types of players in the market (Zhang et al., 2021; Jiang, 2022), or estimate a hierarchical Bayesian η that is internally consistent with the data.

The Nash bargaining solution needs to satisfy equality based on the bargaining power:

$$\frac{\eta}{1-\eta} = \frac{\Delta \pi_{ij(t)}}{\Delta u_{ji(t)}} \tag{24}$$

We can analytically solve for the optimal $\lambda^*_{ij(t)}$ after rearranging the function:

$$\lambda^*_{ij(t)} = \frac{(1-\eta)\beta_i \gamma_{i(t)} + \eta \beta_j \frac{s_{j(t)}}{p_t} + \eta \frac{h_{j(t)}}{p_t} - (1-\eta)\frac{h_{i(t)}}{p_t}}{(1-\eta)\beta_i + \eta \beta_j} \tag{25}$$

There exists a unique solution to maximize joint surplus according to the Nash bargaining. The analytical solution indicates that the discount would be larger if the platform commission rate is higher, the subsidy to the consumer is higher, the hassle of the customer is higher, and the hassle of the driver is lower.

G.2 Estimation and Simulations

We can plug the analytical solution of Equation (25) back into Equation (8) to simulate the choice of leakage. We use the model specification in Section 5.4 for estimation.

$$\begin{aligned}
&Pr[L_t = 1 | \eta, \gamma_{i(t)}, s_{j(t)}, h_{i(t)}, h_{j(t)}] = Pr[\Delta \pi^*_{ij(t)} + \Delta u^*_{ji(t)} + \epsilon_{ij(t)} + \epsilon_{ji(t)} \geq 0] \\
&= Pr[\beta_i \cdot \gamma_{i(t)} p_t - \beta_j \cdot s_{j(t)} - (\beta_i - \beta_j) \cdot \lambda^*_{ij(t)} \cdot p_t - h_{i(t)} - h_{j(t)} + \epsilon_{ij(t)} + \epsilon_{ji(t)} \geq 0] \\
&= Pr[\beta_i \cdot \gamma_{i(t)} p_t - \beta_j \cdot s_{j(t)} - \frac{\beta_i - \beta_j}{(1-\eta)\beta_i + \eta\beta_j} \cdot [(1-\eta)\beta_i \gamma_{i(t)} p_t + \eta \beta_j s_{j(t)} \\
&+ \eta h_{j(t)} - (1-\eta) h_{i(t)}] - h_{i(t)} - h_{j(t)} + \epsilon_{ij(t)} + \epsilon_{ji(t)} \geq 0]
\end{aligned} \tag{26}$$

If $\eta = 1.0$, drivers have the full bargaining power. Drivers will provide a take-it-or-leave-it-offer that makes customers indifferent between off-platform and on-platform transactions.

$$Pr[L_t = 1 | \eta = 1.0] = Pr\left[\beta_i \cdot \left(\gamma_{i(t)} p_t - s_{j(t)}\right) - h_{i(t)} - \frac{\beta_i}{\beta_j} h_{j(t)} + \epsilon_{ij(t)} + \epsilon_{ji(t)} \geq 0\right] \tag{27}$$

Equation (27) shows that the probability of leakage depends on the customers' subsidy and hassle even though customers have no bargaining power (i.e., $\eta = 1.0$). In this extreme case, customers' individual rationality constraint is binding at the equilibrium.

We find that $\eta = 0.975$ yields the maximum log-likelihood in our sample when we use Equation (26) to simulate the choice of leakage. The result suggests that drivers almost own the full bargaining power. It is very likely for drivers to provide a take-it-or-leave-it offer: the discount makes offline transactions weakly preferred by customers.

Given the institutional details, it is not a surprise that drivers have extreme bargaining power. Most drivers are experienced professionals that take jobs regularly. As a result, they have better negotiation skills than customers who are only on the platform for a one-off transaction. Moreover, drivers have complete information about the commission fee the platform charges them, but customers may not know such exact details. However, Table 8 show that experienced users who have repeated transactions on the platform can obtain an offline discount ranged from 16% to 21.6% (see Section 5.3.2). We will discuss what determines bargaining power in Appendix G.3.

The model estimates allow us to simulate the outcomes of counterfactuals by doing a grid of $\eta \in \{0, 0.1, ..., 0.9, 1.0\}$ (see Figure 10) using Equation (26). The worse leakage outcome happens at $\eta = 0.65$ when drivers have slightly stronger power than customers. The simulation shows that the platform would prefer full bargaining power on the driver side ($\eta = 1$) rather than on the customer side ($\eta = 0$).

G.3 More on Bargaining Power

The η is an underlying primitive of a bargaining model. Players with more bargaining power obtain a bigger share of the surplus. This power is captured differently in different contexts. Rubinstein (1982) describes bargaining power as a player's patience (e.g., the discount factor in dynamic models). In other bargaining models (Binmore et al., 1986), bargaining power can represent concepts such as negotiation skills or experience of the player.

The bargaining power is also related to the market thickness for a player (Fong, 2020). Having more available outside options and less competition (Backus et al., 2020) increases the bargaining power. The positive externality of a player might plays a role in determining the bargaining power (Zhang et al., 2021).

With better data, researchers can estimate the bargaining power under different supply-demand relationships, which differs across space, time, and product popularity:

- When the customer has a hard time finding a vehicle, the driver has stronger power.
- When the driver has difficulty getting a job, the customer has high power.
- A scheduled trip on the next day gives a patient customer more time to find a backup driver and thus a potential stronger power in the market.
- A driver with a medium truck is competitive because the vehicle dominates the medium and small van for being able to load more items.
- Experienced drivers or users with repeated transactions on the platform might have better negotiation skills and thus stronger bargaining power.

In our context, the platform can potentially leverage or affect the bargaining power to mitigate leakage by changing market conditions or making strategic matching.