NORTHWESTERN UNIVERSITY

Digitization for Innovative Platform Design and Advertising

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Marketing

By

Yingkang Xie

EVANSTON, ILLINOIS

June 2023

Abstract

Technology that processes text, audio and video, as well as location data, has revolutionized many industries by enabling innovative operations for customer retention. To retain transactions for a platform and viewers for advertisers, this dissertation leverages novel digital tools to analyze consumer behavior, proposes original economic frameworks to guide platform design, and generates new insights to encourage engaging ad creatives.

Chapter 1 and Chapter 2 are devoted to disintermediation, also known as leakage, when users stop doing transactions on the platform. Buyers and sellers can coordinate outside the platform to transact directly, usually to avoid paying fees after being matched. Although platforms are concerned about losing revenue, leakage—by its very nature—is hard to measure and mitigate. Working with Huaiyu Zhu and Jingyi Wang who are employees of an on-demand logistics platform, we develop three detection algorithms and examine their effectiveness in Chapter 1. We find that the state-of-art BERT-based deep learning model improves the precision of detection, but comes with an expensive tradeoff of dropping the high recall we achieve by tracing whether drivers stopped by the origin and destination of a canceled trip. We call for caution regarding heavy resource allocation towards text mining on conversational data when platforms could be more cost-effective by using only behavioral data to detect leakage. In Chapter 2, we focus on the economic incentives behind leakage. 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. We discuss how targeting, monitoring, and matching can mitigate leakage by better aligning the incentives of different parties in two-sided markets.

Chapter 3 (with Joonhyuk Yang, Lakshman Krishnamurthi, and Purushottam Papatla) focuses on ad avoidance when viewers stop watching a video ad. We use a proprietary measure of ad interruption provided by a marketing analytics platform to tracks whether an ad is being played on the screen, and develops a new framework to algorithmically measure the energy level in ad content from the audio of ads. Our machine learning-based measure is related to the human arousal stimulated by ad content. Given that TV ads have become increasingly energetic, we investigate the association between the energy level in ad content and the tendency for consumers to avoid the ads. Overall, more energetic commercials are less likely to be avoided by viewers. However, the association varies across product categories and program genres. We suggest advertisers pay attention to components of ad content other than loudness, which has been regulated by law.

Digital transformation is a mindset, not just about the technology solutions. I hope that these three chapters not only document how novel digital measures can be developed to inform and boost retention efforts for platforms and advertisers, but also demonstrate that marketing researchers and economists can seize opportunities in digitization to work with firms closely to create a culture of innovation and customer-centricity.

Acknowledgements

Pursuing a Ph.D. at Kellogg has been an incredible journey, filled with challenges, discovery, and growth. Completing this dissertation would not have been possible without the support and guidance from my committee - Eric Anderson, Hemant Bhargava, Brett Gordon, Lakshman Krishnamurthi, and Anna Tuchman. I have been very fortunate to work with these world-class researchers. Their expertise in marketing, economics, and technology shapes my research questions and inspires me to think outside the box.

I am indebted to Brett, my committee chair, for his mentorship. Brett's knowledge, foresight, attentiveness, patience, and encouragement made all the difference in my academic career. When tough personal circumstances happened in my life and caused me to be depressed, Brett not only offered sympathy and emotional support, but also helped me create concrete plans to achieve the doctoral milestones required by the department. I appreciate his dedication and care for me while giving me the freedom to deal with life challenges, as well as leaving the country to acquire data, for which I am forever grateful.

I want to thank my four committee members for their positive impact on my research style. Anna and Eric taught me the first quantitative marketing class of my life, in which I learned about field experiments and various applications of causal inference on intriguing marketing topics. Since then, they have been a source of inspiration and motivation for me to conduct in-depth research. Their invaluable feedback has pushed me to think critically and approach problems in new and innovative ways. Hemant and Lakshman

are distinguished experts in their respective fields. They constantly share their wisdom and insights with me, which have greatly improved my understanding of platforms and advertising. I am very lucky to have Anna, Eric, Hemant, and Lakshman on my committee because they always challenge me to delve deeper into my chosen topics and think broader about the managerial implications of my findings. This process has been both exhilarating and humbling, as I always find out how much there is still left to discover.

In addition to my dissertation committee, I would like to extend my deepest gratitude to Joonhyuk Yang and Huaiyu Zhu, my coauthors who are like older brothers from another mother. They have been supportive of my need to travel between the U.S. and China for different projects, and have tolerated my unusual work hours while I am on a different continent from them. Joonhyuk, with who I shared an office and published a paper together, gave me advice ranging from course selection to conference traveling and presentation. Huaiyu, an industry veteran in risk management and on-demand services, never hesitated to share his institutional knowledge and unique insights on how businesses work. Thanks to Huaiyu, I had the opportunity to visit an innovative logistics company and worked with Jingyi Wang to collect and analyze data for exciting research projects. I also want to thank Angela Lee for her kindness and active participation in my academic journey when I need help with preparing talks and writing manuscripts.

Last but not least, I would like to acknowledge my parents for their unconditional love and support. They have always believed in me, whether it was when I decided to study agricultural economics in Kansas, spent an additional year earning a master's degree in statistics at Davis, or quit my Silicon Valley job to pursue a Ph.D. in marketing at Kellogg. I hope to continue making them proud in all of my future endeavors.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Monitoring Disintermediation: Actions Matter More Than Words (joint with Jingyi Wang and Huaiyu Zhu)

1.1. Introduction

Platform-based business models notch up trillions in market capitalization (Cusumano et al., 2020) by helping sellers and buyers find each other and engage in convenient and trustworthy transactions (Einav et al., 2016). They are important to facilitate spot trades in the market for an appropriate charge (Rochet and Tirole, 2006). However, many marketplaces face a problem called disintermediation – suppliers and consumers can transact privately outside the platform to circumvent the marketplace fees. For example, Uber or Lyft drivers ask clients to cancel requests on the marketplace app and pay them offline (Bellotti et al., 2017). A case study at Harvard Business School documents that appromately 90% of transactions at ZBJ.com are conducted outside its freelance platform (Zhu et al., 2018). Firms have incentives to monitor disintermediation to manage revenue leakage and make informed decisions on their pricing and product design.

The first challenge is to detect disintermediation, which is hard to monitor given the nature of offline coordination. Although firms (e.g., Uber, Airbnb and eBay) have official policies to prohibit off-platform transactions with threats on account suspension (Uber, 2022; Airbnb, 2022a; eBay, 2022b), it is unclear whether they have ways to identify this fraudulent behavior and whether their detection technology is cost-effective. Firms

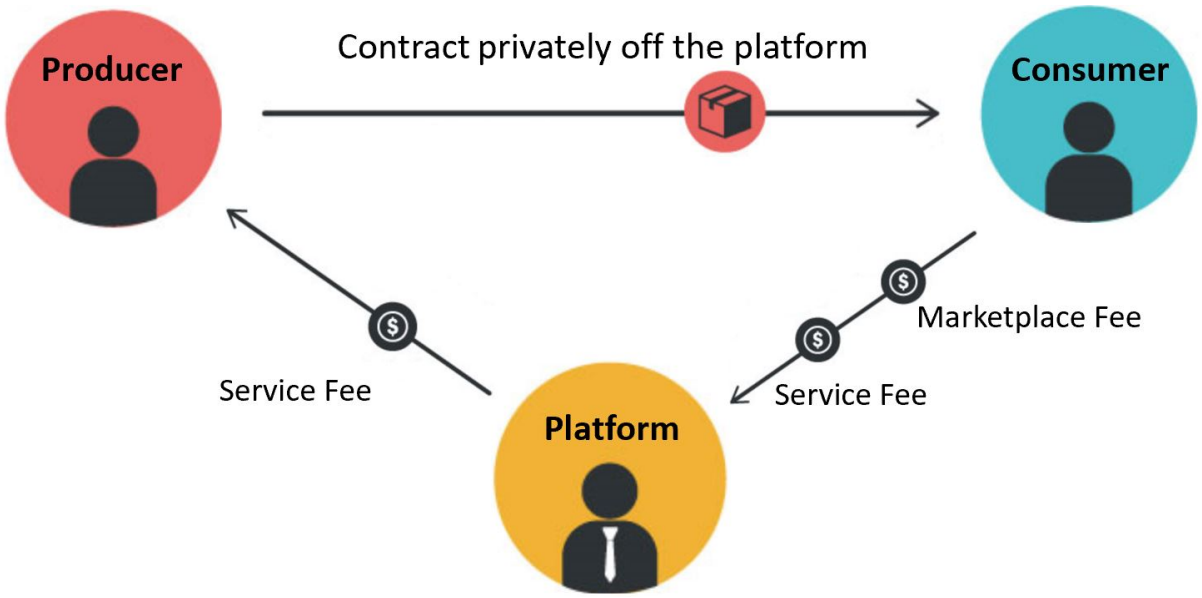


Figure 1.1. Off-platform Transaction to Avoid Marketplace Fee

face many choices in leakage detection: 1. What data should be collected (e.g., text vs. behavior)? 2. What algorithms should be used (e.g., rule-based vs. deep learning)? 3. What is the more sustainable solution given stricter privacy regulations in the world?

To answer the above questions, we take advantage of the geolocation (e.g., GPS footprints) data and the in-app conversations (e.g., voice calls and instant messages) provided by Lalamove, which owns the largest on-demand logistics platform in China and operates internationally across Latin America and North America. Typically, drivers create digital data in space and conversations with the following timeline:

(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).

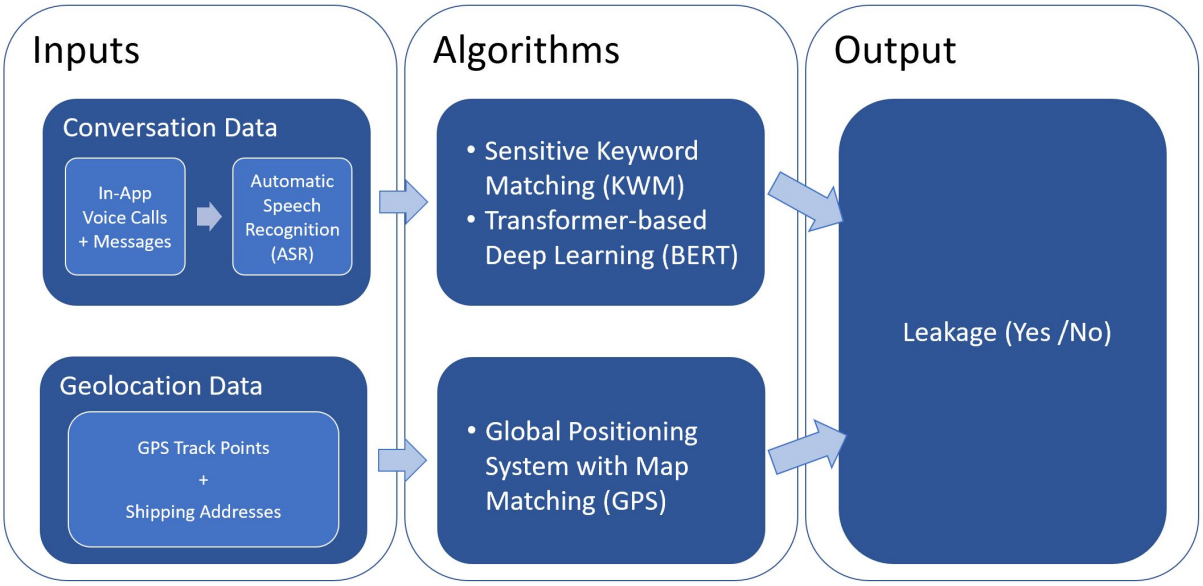


Figure 1.2. The Automated Monitoring Framework of Leakage

(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.

We develop and tune algorithms with a large set of past transaction data after iterations of offline evaluation. The detection solutions are then deployed in the real-world production environment with structured query language (SQL) and natural language processing (NLP) transformer models BERT (Devlin et al., 2018). Figure 1.2 is the flowchart of data and algorithms in our fraud detection system. We evaluate the performance of these solutions using a new dataset manually classified by human experts. By reviewing the recorded conversations, drivers' movements, and other transaction information, these experts confirmed 1009 disintermediated transactions out of randomly selected 5557 cancellations without knowing any algorithmic predictions (see Section 1.4.1).

Many firms (e.g., Airbnb, eBay, Didi, ZBJ) monitor conversations for leakage detection and safety protection (Airbnb, 2022b; eBay, 2022a; Shen, 2019; Gu and Zhu, 2021). Although conversation data cover 85.5% of the cancellations and provide very rich information in our application, we find that matching the keywords and phone numbers can only recall 58.67% of the disintermediated transactions identified by human experts. In contrast, our deployed system using behavioral data – whether drivers passed by the origin and destination of a previously canceled transaction around the time of service – are more cost-effective and able to recall 94.05% of leakage. The significant gap in the recall ratio (58.67% vs. 94.05%) demonstrates the problem of text mining solutions in leakage detection, because trading partners 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). Without using behavioral data as an additional source of information, firms that only rely on conversation data could underestimate the scale of disintermediation. In a production environment, we find that the keyword matching and state-of-art BERT-based deep learning models have 7 times and 14 times as many false negatives as the GPS-based system, respectively.

Our results provide broad implications for data collection and user privacy. First, we call for caution on heavy resource allocation toward active listening to users' conversations because of its ineffectiveness in fraud detection. Although monitoring conversations is a common and established practice (Airbnb, 2022b; eBay, 2022a; Gu and Zhu, 2021) for leakage detection in the industry, we find that our deployed system can achieve a higher F1 score without using the conversation data. Online marketplaces that feature location-based services (LBS) do not need to collect more data to prevent revenue leakage.

Secondly and more importantly, the overall trend in privacy regulation may limit the platform's ability to capture conversational data continuously. We suggest that engineers and data scientists seek and create products that motivate user-consented data. In our case, geolocation data are voluntarily shared by drivers in exchange for job assignments because they are actively "listening" to nearby matches as job seekers rather than being "listened" to without consent.

To the best of our knowledge, our work is the first study that develops and compares detection solutions on disintermediation. We demonstrate how to use voice and geolocation data and discuss the tradeoff between precision and recall in fraud detection. Our findings also contribute to NLP applications by calling for actions to obtain behavioral data to help evaluate text mining solutions. Engineers and data scientists may test their algorithm with only text data they use for training, thus misreporting the recall without leveraging behavioral data that can improve the data labeling quality. A leadership role requires one to engage in strategic thinking on how to examine the actual performance of a system and determine what data mining solutions are more cost-effective and less privacy intrusive to achieve the business goal. While actively listening to users' conversations via phone or smart devices is an emerging domain in the Internet of Things (IoT), we call for caution on using such technology to fight fraud or crime not only because of the privacy controversy but also its ineffectiveness.

1.2. Motivation

1.2.1. The Damage of Disintermediation

One key consideration for the success of platform businesses, particularly online marketplaces, is the extent of the leakage problem and how they may deal with it (Hagiu and Wright, 2021). Disintermediation is the Achilles heel of the platform business model: leakage is a flaw that is hard to measure but undermines the platform's ability to continuously generate revenue from the value that it created (Ladd, 2021). The gradual and subtle process turns out to be a money leak for platforms, so practitioners also call disintermediation as platform leakage. Platforms need technology to detect the gradual and subtle process of leakage.

The requirement of individual interaction is one reason why sharing economies and gig economies (Madden, 2015; Said, 2015; Sarva and Wald, 2015; Zhu et al., 2018) might have been hard hit by disintermediation. The delivery of the services or products is handled directly by the seller, so as the communication with buyers. In such cases, leakage is likely to happen as the trading partners can share information, gain trust, and make direct payments. Marketplaces with these features are perfect environments for collusion without the platform's consent. While people think that cutting out the middleman only happens after the completion of the first job, we want to point out that leakage can occur instantly at the moment when the trading partners find each other – the platform becomes a showroom (Wang and Wright, 2020).

Besides online labor marketplaces, it is not uncommon to observe leakage in other industries. Retailers, travel agencies, financial brokers, real estate agents, and advertising

networks more or less encounter disintermediation as long as transactions can proceed with less cost or effort by cutting out the middleman. Some intermediaries (e.g., Amazon and Expedia) are claimed to be protected by minimum advertised pricing (MAP) to avoid being the showrooms for suppliers. However, these platforms still constantly check supplier and competitor prices to forestall disintermediation. We can think of MAP monitoring and enforcement as an alternative strategy to leakage detection and deterrence.

In the context of Lalamove (think Uber for cargo vans and trucks), drivers ask customers to cancel the delivery request after being matched on the platform. We observe cases where drivers propose off-platform coordination using WeChat for communication and payment. Lalamove are concerned about losing their cut, which is an explicit threat to their revenue. Moreover, Lalamove wants to investigate the extent of the leakage problem to discover any implicit problems such as overpricing or pain points in platform services that discourage consumers from finishing the transaction on the platform. Lastly, the absence of protection and post-transaction review might result in poor service quality or safety issues. It is a priority for Lalamove to retain and protect transactions on the platform to forestall reputational damage on the brand and other drivers. Managing revenue losses and potential reputational risk requires the right technology for leakage detection and deterrence.

1.2.2. The Limitation of Related Work

Many online marketplaces use the user-generated text for leakage detection and deterrence. For example, Airbnb (Airbnb, 2022b) scans and analyzes messages to protect users from being at risk for communicating or paying outside the platform. When users and

hosts exchange contact information via the messaging service on the platform, their phone numbers or external URL links in conversations are hidden by Airbnb. eBay (eBay, 2022a) replaces email addresses with aliases to hide personal contact information and states that they monitor messages (together with any attachments) sent between members for policy violations. A freelancer market (similar to Upwork) (Gu and Zhu, 2021) refined its algorithms and dictionaries over years to detect and deter disintermediation with sensitive words. Platforms in the gig economy, sharing economy, and e-commerce can benefit from cutting-edge NLP methods to improve detection precision to a large extent.

However, it is unclear whether conversation data alone can help firms recall most fraudulent cases. There are unknown amounts of false negatives (i.e., true leakages that are not flagged) due to the alternative channels to coordinate off the platform (e.g., chat in person or connect on social networks). Users who know they are monitored can also encrypt their contact exchanges (e.g., add random characters or sounds in between phone numbers) or communicate in alternative formats[1] (e.g., an image with email address) to hide their fraudulent activities. Therefore, firms may inappropriately evaluate the performance of their detection solutions when engineers and data scientists limit themselves to conversation data without using behavioral data for additional validation.

To evaluate the performance of conversation data and improve the existing practice, we take advantage of GPS track points and transaction-level information provided by Lalamove. To benchmark on industry standards, we implement both the simplest and cutting-edge NLP models. Thanks to novel data and algorithms, we now have the unique

[1]Vacation lords discuss ways to share email address with consumers on Airbnb: https://www.vacationlord.com/how-to-get-around-airbnbs-email-blocking/

opportunity to compare leakage detection solutions with discussion on the cost and benefits of unstructured data (e.g., conversation) versus structured data (e.g., geolocation) for different fraud detection[2] purposes.

Table 1.1. Fraud Detection Solutions for Platform Leakage

Data	System	Algorithm
Conversation	Rule-based	(a) Drivers are involved in conversations with sensitive keywords on exchanging contact information (e.g., phone number) or on proposing an offline coordination (e.g., avoid the platform)
Conversation	Deep Learning	(b) Drivers are involved in conversations that are flagged by the BERT-based NLP model, which was trained by past cancellation data labeled by analysts and customer representatives
Geolocation	Rule-based	(c) Drivers have at least 15 track points each near the origin and destination of the canceled request around the time of the service, and shippers don't have similar requests that are fulfilled on the platform

Notes: To maintain the proprietary nature of the fraud detection algorithms, we do not disclose the specific values of time and distance criteria to avoid the information being used by drivers or shippers to game against the systems.

1.3. Detection Algorithms

1.3.1. Target Variable

The task is to construct a binary classifier to predict whether a transaction was disintermediated. The on-demand logistics platform provides us with two types of data to build the fraud detection systems for disintermediation: (1) conversation data generated from

[2]The company considers a disintermediated transaction to be fraudulent because striking private deals go against the terms of service, leaving the transaction unprotected, and may involve intentional deception to secure an unfair gain.

conversations between drivers and customers via the in-app voice calls and instant messages; (2) geolocation data generated from the Global Positioning System on the drivers' devices. Fraud detection solutions are the products of data and algorithms (see Table 1.1).

Our target variable is instant leakage rather than subsequent leakage in this investigation. Instant leakage includes the pre-specified delivery request being leaked immediately after the matching. Subsequent leakage represents all recurring businesses within the pair of driver and customer where these future delivery requests may not be initiated on the platform. For traceability, we focus on instant leakage because a canceled request provides the shipping addresses and time of the service as the starting point to investigate whether a specific demand is leaked. We assume that customers provide real and accurate shipping addresses in the request.

1.3.2. Data

Lalamove matches drivers and customers for moving goods and furniture. The platform makes more than one million matches every day with more than 20% of them being canceled. The manual labeling process is costly given the large volume of cancellations. We want scalable detection solutions that are on par with human.

This section will describe the conversation and geolocation data and how they are generated. All in-app conversations are linked to delivery requests: drivers and customers can call or write a message to each other on the platform. GPS track points are uploaded by drivers' cell phones when the app is active or running in the background after obtaining user consent.

1.3.2.1. Conversation Data (Unstructured). When a driver or customer contacts each other regarding a shipment request, the call or message is first connected to the anonymization server to ensure that personal contact details are protected, preventing trading partners from being spammed or harassed. The anonymization server stores all the audio and text generated from the conversations. Additionally, an automatic speech recognition (ASR) service converts audio into text in preparation for text mining tools.

Conversations are linked to delivery requests. Regarding text coverage, 85.5% of canceled requests have conversational content, which indicates that most drivers and customers would communicate with each other before they back out of the match or proceed with off-platform coordination. Among these cancellations, 82.9% of conversational content is generated by voice call only, 15.7% are contributed by the usage of both voice call and instant messages, and 1.4% are from messages only.

The depth of information in the text provides us with verbal descriptions of why customers or drivers cancel the request. The data labeling team finds that 70.7% of cancellations conversations contain useful information to classify the stated reasons behind cancellations. The platform observes 61.4% of cancellations are non-leakage related and 38.6% are leakage related. The most common non-leakage related reasons are the mismatch of vehicle and shipping items (51.4%), the change of driver's availability (12.6%), customer's cancellation (12.0%), while the most common leakage-related reasons are incentive conflicts on fees (28%), the change of time of service (24%) and shipping addresses (12%), agreement of new price after negotiation (7%). The leaked transaction is not only motivated by the opportunism to avoid intermediary's fee as previously mentioned in the literature (Gu and Zhu, 2021; Ladd, 2021), but could also be triggered by the need to

modify other contract terms such as service details and the corresponding payment. Platforms could use text information to identify the pain-points of on-platform or off-platform transactions to help improve their products.

This labeling process may not correctly reflect the actual percentage of leakage and the proportion of different motivations behind leakage – both drivers and customers can hide or lie about their real reasons for their cancellation in their phone or text conversations. However, we can still learn from the rich text data that a leaked transaction is not only motivated by the opportunism to avoid intermediary's fee as previously mentioned in the literature (Gu and Zhu, 2021; Ladd, 2021), but could also be triggered by the need to modify other contract terms such as service details and the corresponding payment. Platforms could use text information to identify the frictions of on-platform or off-platform transactions to help improve their products.

1.3.2.2. Geolocation Data (Structured). GPS track points are uploaded when a driver uses the platform app on their cell phones or has the app service running in the background after obtaining user consent. Presumably, geolocation data is complete for drivers looking for the next job regardless of drivers being idle or occupied by an existing job. The platform broadcasts a 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.

Raw GPS track points consist of longitude, latitude, and timestamp, which are all attached to drivers. Connecting the points with lines can form a track that represents the path of drivers. The track points allow the algorithms to calculate the distance between the shipping addresses (i.e., origin and destination) and the drivers' locations to determine

whether a driver is near the points of interest. We can also infer the time drivers spend at a location and whether they are stationary or moving.

1.3.3. Algorithms

Table 1.1 describes how the systems use the text and geolocation data to detect leakage. We convert all conversational data from audio to text in preparation for NLP methods and link all related GPS track points to a canceled request based on the shipping addresses and the time of service in preparation for geolocation detection rules. For rule-based systems, we want to use algorithms that reach high recall or high precision to find performance boundaries in our empirical evaluation. For geolocation detection rules, we use heuristics based on insights from human experts who manually classify fraudulent transactions in the past. All the algorithms developed or trained offline will be implemented in a production environment and compared to human classification.

1.3.3.1. Rule-based Text Mining - Keyword Matching (KWM). One way to perform a simple NLP task is to conduct vocabulary and phrase matching. It is common for firms to check if users exchange contact information (e.g., Airbnb, eBay). In the on-demand logistics platform, the most common communication tools on the go are WeChat and cell phones, so we use these tool names to infer a potential contact exchange. Moreover, sensitive words listed by Gu and Zhu (2021), which implies the proposal of transacting outside the platform or describes the motivation of avoiding fees, are helpful to detect leakage. 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 on WeChat. In summary, we perform simple "OR"

statements of sensitive keywords and linguistic patterns using the regular expression to classify cancellations. Examples include:

- Keywords of contact exchanges: "cell", "phone", "wechat"
- Keywords of leakage proposal: "avoid platform", "private deal", "offline transaction", "information fee", "service charge"
- Consecutive phone numbers: "158" or "one five eight".

With the broad text match, we expect a low precision given possibly many false positives (i.e., innocent cancellations that are falsely classified as leakage). Nonetheless, our goal for this algorithm is to include all relevant hints of leakage to achieve high recall.

1.3.3.2. Deep Learning Text Mining - Bidirectional Encoder Representations from Transformers (BERT). BERT has delivered state-of-the-art performance on many NLP tasks (Devlin et al., 2019). The method trains a general-purpose "language understanding" model on a large text corpus (e.g., users' conversations) and then uses that model for downstream NLP tasks (e.g., leakage detection). It outperforms the traditional single word embedding representation for each word in the vocabulary because it has bidirectional contextual representations of each word based on the other words in the sentence.

We used the pre-trained BERT-Base Chinese model on TensorFlow for character-based tokenization. The masked-language model was already trained by Google Research with 12-layer, 768-hidden, 12-heads, 110M parameters on Chinese simplified and traditional Wikipedia. Conditional on the pre-trained transformers, we trained the softmax classifier using past canceled delivery requests with the binary label of leakage generated by analysts and customer representatives. We preprocessed the input text files by removing modals

and duplicate words after using the automatic speech recognition (ASR) technology to convert the voice calls.

We choose BERT as the state-of-the-art benchmark because it has become a ubiquitous baseline in NLP experiments (Rogers et al., 2020). The max input length was set to 256, and the training batch size was 32 to achieve acceptable computation time under the constraint of GPU memory. Other parameters such as learning rate are tuned through cross-validation.

1.3.3.3. Rule-based Geolocation Mining - Global Positioning System (GPS). The algorithm flags a cancellation as a leakage if the driver passes by the origin and destination around the time of service of a previously canceled order. This tracking of geolocation provides us with a new dimension of information to trace leakage. Our algorithm calculates the distance between drivers' GPS track points and the shipping addresses to check whether the driver is present within a small radius of both the loading and unloading locations specified by the customer[3]. The small radius is tuned through offline iterative process to achieve good F1 score before the algorithm is officially deployed in production.

To reduce false positives in which drivers pass by the points of interest coincidentally without a stop, we add a restriction to require at least 15 track points each near the origin and destination to classify the cancellation as leakage because an offline transaction would need sufficient time to load and unload the shipping items. The number of minimum 15 track points is discovered through previous iterations of algorithm evaluation with a grid

[3]We assume that customers are truthful on shipping addresses to guarantee the effectiveness of geolocation data in recalling leakage. Customers sometime change their shipping addresses, but most of these new addresses are not far away from the original ones.

search. Moreover, the rules also check whether customers have other complete transactions with a similar trip distance around the time of service. This additional rule prevents us from classifying a previously canceled order as leakage which turns out to be fulfilled on the platform in another delivery request with the same or a different driver.

We develop rule-based geolocation detection to ensure transparency and explainability. While machine learning models such as random forests or xgboost could improve classification performance, it is important for us to explain how the algorithm make a decision if the platform decide to warn or punish cheaters.

1.4. Empirical Evaluation

We evaluate the three solutions depicted in Table 1 and their combinations (i.e., voting ensemble of multiple models). We want to compare the geolocation solution to the common monitoring solution using conversation data with the state-of-the-art NLP benchmark. The following setup allows us to report post-launch performance:

(1) Algorithms automatically label each transaction (see Figure 1.2);

(2) Human experts independently label cancellations without knowing the algorithms;

(3) We use human labels as ground truth to evaluate the performance of algorithms.

1.4.1. Human Labeling

Manually labeling all cancellations is costly given the large number of transactions (i.e., one million matches per day). Due to the labor constraint, human experts labeled a randomly sampled 5557 cancellations and confirmed 1009 disintermediated transactions.

To be qualified as a suspicious cancellation, the customer and the driver are involved in an on-platform conversation with at least one suspicious keyword (e.g., cell, phone, wechat) for exchanging contact information before the cancellation of the shipping request, or the assigned driver has at least one GPS track point each near the origin and destination (i.e., the loading and unloading addresses) of the previously canceled job after the shipping request is canceled. This pre-filtering allows us to collect more relevant positive instances to compare mining solutions by investing the same amount of labor hours in the investigation. After all, human experts will not flag a cancellation without seeing evidence of contact exchanges or track points near the points of interest.

Human experts only classified cancellations because a complete trip is not leakage by definition in our context. For example, a customer initiates a shipping request which is then assigned to a driver by the platform. After communication, the customer and the driver agree to take this specific request outside the platform. They will need to cancel the active request on the platform to proceed with offline coordination without making a payment via the platform. This is instant leakage, with the current demand being leaked immediately. In another context, subsequent leakage could exist when the customer and driver potentially trade in the future for recurring business after establishing a relationship. We do not ask our human experts to label the subsequent leakage in this investigation because we cannot verify whether a future demand exists. We can link the instant leakage, not the subsequent leakage, to pre-specified requests with information about the time of service and shipping addresses. For traceability, we thus focus on instant leakage rather than subsequent leakage.

To identify a disintermediated transaction, human experts review the activities in cancellation to check whether the driver and customer contact on the platform using in-app voice calling or instant messaging. If so, they will listen to the call conversations and read text messages to determine if anyone proposes leakage, and judge whether the driver and customer reach agreement to transact outside the platform. They will also check whether the driver passed by the origin and destination around the time of service of the canceled order. When there is any evidence of leakage, human experts check whether the customer files other requests that share similar shipping addresses specified in the previously canceled order. Any information about new addresses or the time of service mentioned in the conversations could be used to prove or disprove the existence of leakage. In addition, human experts will make a judgment on whether the driver stays at the points of interest with sufficient time to load or unload items. According to human experts, they will make exceptions if the driver fulfilled another request on the platform with a similar origin and destination on the same day. Although cancellations are evaluated only once by the human experts in the labeling team, all instances are re-evaluated by data analysts in another team with no disagreement.

Admittedly, human experts could misclassify leakage when data is incomplete or incorrect. For example, drivers could turn off their GPS or completely shut down the phone to avoid being tracked by the platform. Poor cell phone signals might also affect the trackpoint uploads of some drivers. However, this experiment focus on evaluating whether our scalable algorithmic solutions can deliver performance that is on par with human experts. Therefore, we treat human labels as ground truth.

1.4.2. Evaluation Metrics

Using human-tagged labels as the ground truth, we evaluate the recall and precision of the algorithms (see Table 1.2). These metrics could be calculated based on the confusion matrix that tabularizes the predictions and actuals in a contingency table. The columns of the confusion matrix represent the actual class (i.e., human-labeled leakage), and the rows of the table indicate the predicted class (i.e., system-predicted leakage). The confusion matrix demonstrates four terms: True Positive (TP), True Negative(TN), False Positive(FP), False Negative(FN). Precision is the ratio of correctly predicted leakage to total predicted leakage. Recall is the ratio of correctly predicted leakage to all observations of the actual leakage. F1-score is the harmonic mean of precision and recall.

Table 1.2. The Evaluation of Detection Solutions

	Label:Yes	Label:No	Precision	Recall
Model:Yes	TP	FP	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$
Model:No	FN	TN		

Note: TPs are hits, FNs are misses, FPs are false alarms, and TNs are correct rejects.

1.4.3. Empirical Results

Table A.1 demonstrates the post-launch performance of the three leakage detection solutions and their combinations. We want to compare our novel solution using geolocation data with the common monitoring solution using conversation data processed by the state-of-the-art NLP model. We deployed the algorithms in a production environment before the evaluation to ensure scientific objectivity. Human experts classify cancellations without knowing the algorithmic predictions in the production.

Our human experts found 1009 disintermediated transactions in 5557 cancellations. Using their labels as the ground truth, we find that the common industry practice of monitoring conversations between trading partners recalls less than 58.67% of collusion when we use keyword matching (KWM) that indicate the exchange of contact information or the proposal of offline coordination. In contrast, geolocation data, which trace drivers' behavior using the global positioning system (GPS) given a canceled request with the specified origin and destination of a trip, can recover up to 94.05% of leakage. Tracking drivers' GPS footprints near the origin and destination results in a much higher precision of 90.38% than the 38.00% delivered by using the conversation data only. While deep learning with text mining - Bidirectional Encoder Representations from Transformers (BERT)- achieves a precision of 94.52%[4], at the cost of a lower recall ratio of 13.68%.

Table 1.3. The Performance of Detection Solutions

Data	Model	Classification	Label:Yes	Label:No	Precision	Recall	F1-Score
Conversation	KWM	Model:Yes	592	966	38.00%	58.67%	0.461
		Model:No	417	3582			
Conversation	BERT	Model:Yes	138	8	94.52%	13.68%	0.239
		Model:No	871	4540			
Geolocation	GPS	Model:Yes	949	101	90.38%	94.05%	0.922
		Model:No	60	4447			
Both	GPS&BERT	Model: Yes	136	7	95.77%	13.48%	0.236
		Model:No	873	4542			
Both	GPS|KWM	Model:Yes	1009	1048	49.05%	100.00%	0.658
		Model:No	0	3500			

[4]To make a fair comparison, we apply different thresholds on the BERT predictions to determine whether a cancellation is a disintermediated transaction. One threshold results in a similar precision of 90.00% as the GPS algorithm but comes with recall of 14.27%. Therefore, GPS algorithm can reach far better recall with the same precision.

The false positives from using the KWM and BERT solutions are 7 times and 14 times as many false negatives as compared to using the GPS solutions. Our simple rule-based GPS detection solution not only achieves a good precision above 90% but more importantly recalls most off-platform transactions, dramatically improving the F1-score compared to the benchmark text mining solutions.

Next, we explore the combination of conversation and geolocation data in leakage detection. We report the two combinations that focus either on precision or recall, which illustrate two strategies for leakage detection and deterrence. One combination is to determine leakage when a cancellation is flagged by both the GPS and BERT solutions (GPS & BERT). Using both geolocation and conversation data provides solid evidence of a delivery request being leaked by the driver and the customer involved in a cancellation. Another combination is to flag cancellations as leakage when flagged either by the GPS solution or the KWM solution (GPS | KWM). The geolocation and conversation data can complement each other in recalling all fraud cases human experts labeled.

1.4.4. Precision vs. Recall

Firms can trade off recall against precision for different purposes. One use case is to generate a "short list" from the system (GPS & BERT) with a very high precision. The short list help prioritize efforts in the investigation of individual cases for warning and lecturing. The firm can also use the predictions to allocate marketing actions. For example, if the estimated probability of the fraud is high, the firm could provide additional incentives to keep the transaction on the platform (e.g., lower marketplace fees). Our results show that 90.38% of leakage flagged by the GPS-based system are true disintermediated

transactions. With additional verification from the BERT-based system, the precision can achieve a 95.77% precision. Although the recall drops significantly from 94.05% to 13.48%, this combination can help firm operations for treating drivers with punishment or retention offers.

Another use-case is to create a "watch list" from the system system (GPS | KWM) by pushing the boundary of recall, conditional on a satisfactory precision. This extensive list includes as many fraudulent cases as possible to track the overall level of leakage. The aggregate number helps firms to evaluate whether a policy change, product launch, or pricing adjustment might increase or decrease the overall leakage. Without the metric, a platform could underestimate the extent of leakage and make uninformed decisions like overpricing. Lastly, repeated observations of the same individual in this "watch list" provide an alternative strategy to confirm cheaters who are recidivists that escape the "short list".

1.4.5. Discussion

Many platforms actively listen to users' conversations for leakage detection and deterrence. For example, the short-stay rental platform Airbnb and e-commerce platform eBay scan and hide phone numbers, websites, or email addresses in their messaging service. The ride-sharing platform Didi Chuxing records calls and in-vehicle conversations. An online freelance marketplace uses sensitive keywords to monitor leakage from conversations on its platform.

Despite being a common practice in the industry, an automatic system of monitoring conversations may fail to achieve a good recall ratio because cheaters can find ways to encrypt their contact exchanges (e.g., send an image with email addresses, add meaningless characters or sounds in conversations) or communicate via other channels (e.g., chat in person or connect on social networks). All these strategic behaviors of cheaters increase the cost of leakage detection for firms. Many cutting-edge NLP methods could improve the precision of the fraud detection system. However, the actual recall may fall out of the radar in the process of algorithm evaluation. Machine learning experts with skillful text mining techniques might limit themselves to conversation data without actively looking for behavioral data for additional verification.

We suggest that leadership in data science to be creative in seeking behavioral data to validate text mining solutions or develop behavior-based solutions. Examples include (1) geolocation data that could be leveraged by firms such as Lalamove (whether driver passes by the shipping addresses of a canceled request) and Airbnb (whether the traveler opens the app near the vacation rental of a previous inquiry), and (2) vacancy data that could be used to determine the availability of suppliers or the consumers, respectively.

We find that GPS footprints is more cost-effective in leakage detection than monitoring conversations. Geolocation data is continuously collected by the on-demand logistics firm and could be efficiently utilized with simple SQL queries. Rule-based GPS detection system could achieve good precision with much higher recall than the systems powered by text mining solutions. Although the BERT-based deep learning approach can achieve high precision, it is computationally expensive to process conversation data with automatic speech recognition (ASR) and natural language processing (NLP). Nonetheless,

conversation data provides the opportunity to help firms understand the motivations behind disintermediation. As demonstrated in Table 3, geolocation and conversation data complement each other with different merits to improve precision and recall when they are in combination. Behavioral tracking is helpful for fraud discovery, and text mining provides additional proof for affirming fraud. Lastly, the firm is actively promoting IoT devices on vans and trucks to help protect customers' safety and assets. Although installing physical monitoring devices incur expenses, in-vehicle tracking provides 24/7 geolocation data with video recordings that are highly reliable behavioral data for fraud detection and deterrence.

Our systems help the firm to prioritize efforts in investigating fraud. We mark the transactions that are highly likely to be related to the leakage, and the firm can investigate those cases with higher priority with potential treatments such as warnings, lecturing, or punishment. The firm can also use predictions to allocate marketing actions. For example, if the estimated probability of the fraud is high, the firm could provide additional incentives to keep the transaction on the platform (e.g., lower marketplace fees). Nonetheless, it is also essential to track the aggregate number of disintermediated transactions that covers most fraudulent cases. The metric can help firms examine whether a policy change, product launch, or pricing adjustment might increase or decrease the overall leakage rate. Without detection technologies with high recall and acceptable precision, firms could underestimate the scale of leakage and make uninformed decisions (e.g., overpricing in commission) without noticing revenue losses.

1.4.6. Limitations

Some disintermediated transactions may pass the human examination because human judgments are susceptible to data availability and correctness. For example, drivers could turn off their GPS or completely shut down the phone to avoid being tracked by the platform. Poor cell phone signals might also affect the trackpoint uploads of some drivers. To overcome these challenges, Lalamove is experimenting with installing IoT devices on vehicles to obtain better coverage of geolocation data. Moreover, the effectiveness of geolocation data in leakage detection also relies on the customers being truthful and consistent on shipping addresses. It is unlikely that customers will provide false addresses on purpose to cover up the real destinations of drivers before being matched, but, in very rare cases, they might modify their demand with new shipping addresses that are far away from the original ones.

Although the deployed GPS solution achieves desirable performance for leakage detection in on-demand logistics, we will need to customize the algorithm and data collection process for other shipping services such as less-than-load (LTL) that features shared truckloads (i.e., freight carpool). Multi-stop trips with large shipments and long-distance deliveries often span a long period of time, which poses a challenge to tracking leakage. Location-based services such as Didi Chuxing, Uber, Wingz could use similar solutions to detect leakage; however, platforms in other contexts may need creative solutions to mine the transaction and geolocation data. For example, Airbnb can check whether a traveler is near the vacation rental of a previous inquiry and whether the vacation rental is not available during the same period of time.

While our detection solutions focus on instant leakage, future research can extend our work to manage subsequent leakage. There exist recurring businesses between producers and consumers after establishing a relationship. The detection of subsequent leakage requires systems to leverage the information from longitudinal data (or panel data) to predict and verify whether repeated demand occurs and whether the future demand is fulfilled outside the platform. No platforms want to see that their businesses are gone after making a match. Firms need new tools for knowledge discovery in databases to make informed decisions about their pricing structure for repeated transactions and how much to invest in customer retention and product development.

1.5. Conclusion

Two-sided markets match buyers and sellers to help facilitate transactions for an appropriate charge (Rochet and Tirole, 2006). Unfortunately, these matchmakers suffer from disintermediation when the two sides of the market collude to reduce their cost by cutting out the middleman. The nature of offline coordination makes it hard to tell whether a canceled transaction is disintermediated. Using novel data provided by the on-demand logistics platform, we demonstrate the effectiveness of geolocation and conversation data in tracking leakage.

While many platforms actively listen to users' conversations and invest in cutting-edge NLP methods, we call for actions to obtain behavioral data to evaluate the cost-effectiveness of their text mining solutions. Our contribution is to show that such common industry practice results in many false negatives (i.e., failures to recall true leakage). As a result, platforms could significantly underestimate the number of disintermediated

transactions. Specifically, the keyword matching and BERT-based deep learning solutions have 7 times and 14 times as many false negatives as our GPS solutions, respectively. We find that conversation data recall less than 60% of the disintermediated transactions when we include as many sensitive keywords as possible. In contrast, the tracking of drivers' behavior can recall almost 95% of leakage with a high precision and explainable rules. We expect further performance improvement by applying machine learning to learn heterogeneous rules for different cities, businesses, and driver types.

Our detection solutions can be generalized to online marketplaces for location-based services (e.g., Didi, Dolly, GoShare, Lyft, Uber, Wingz). However, platforms without constant location tracking may need different solutions to mine their user-consent data in a creative way. For example, Airbnb can check whether a traveler opens the app near the vacation rental of a previous inquiry conditional on its vacancy information. While the current solutions focus on instant leakage, future research can extend our work to measure and manage subsequent leakage, which are the recurring offline businesses between producers and consumers. The detection of subsequent leakage requires platforms to leverage longitudinal data to predict and verify whether repeated demand occurs and whether the future demand is fulfilled outside the platform.

Our results echo Google's Research Director Peter Norvig: "More data beats clever algorithms, but better data beats more data." Firms may not always benefit from collecting "more data" for marginal improvement in algorithms by triggering privacy concerns. We need to act intelligently on deciding what is the "better data" for different applications. One direction is to look for behavioral data explicitly known to be collected (e.g., drivers would not get jobs nearby if they refuse to disclose their current locations) and with a

hassle to hide (e.g., the inattention to turn off GPS or inconvenience for shutting down the phone). Another direction is to look at behavioral changes triggered by experimental variation in the incentives (e.g., fee, subsidy, new product features). Contextualization with interactions between behavioral and conversation data is also worth exploring because the same set of words may manifest different intentions at different locations.

One limitation of fraud detection is that data could become less reliable over time due to the strategic behaviors of cheaters who know they are watched and are threatened by punishments. Since the game of inspection and cheaters' countermeasures is an inevitable race, practitioners and researchers can think ahead about the cost-effectiveness and privacy concerns of different detection solutions. We want minimal data conditional on user consent to achieve our business goals. Behavioral data may be a more reliable way to detect fraud when the cost for cheaters to hide their intention in actions is higher than in words. After all, it is actions, not mere words, that matter.

CHAPTER 2

Platform Leakage: Incentive Conflicts in Two-Sided Markets (joint with Huaiyu Zhu)

2.1. 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, 2023; Ladd, 2021). 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

[1]Theorists study disintermediation with analyses on leakage in online marketplaces (Chaves, 2018; He et al., 2020; Hagiu and Wright, 2023; 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)

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 more 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 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, 2022b). 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

[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.

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 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,

[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 2.3.3 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.

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 parties 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).

[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 A.7 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).

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 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, 2023). 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)

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

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.

2.2. Empirical Setting

Disintermediation undermines the platform's ability to capture the value it creates (Ladd, 2021). 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.

2.2.1. 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 2.1 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 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

[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.1 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.

canceled by the customer without any penalty. The job will be marked as complete after the delivery.

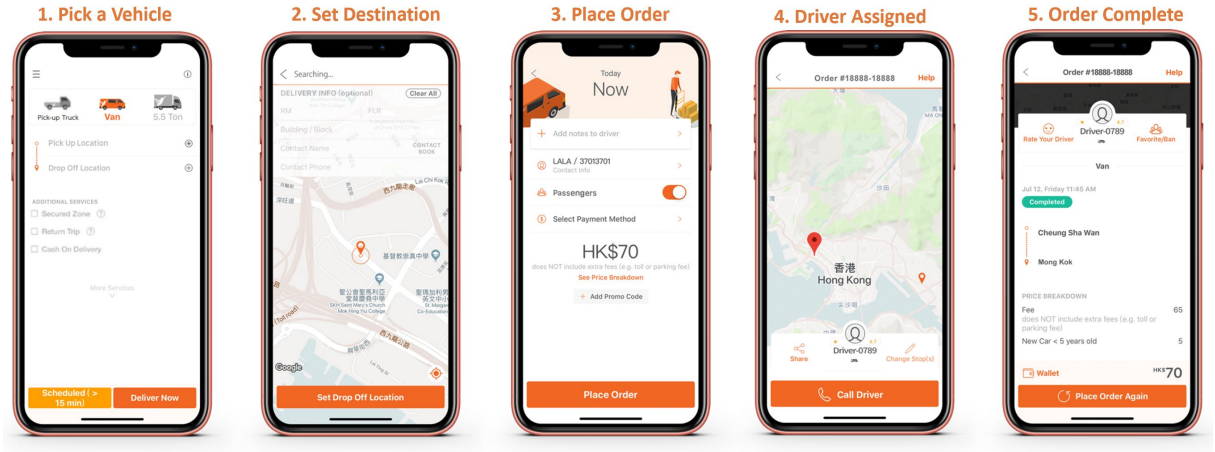


Figure 2.1. The Order Transaction Flow for Cargo Delivery in the App

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.

[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.

This paper uses data in 2019 before the COVID-19 pandemic. During the staggered 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 2.1) to use the free service as Non-VIPs or pay for the premium service as Super-VIPs:

Table 2.1. 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.

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

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.

2.2.2. 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.

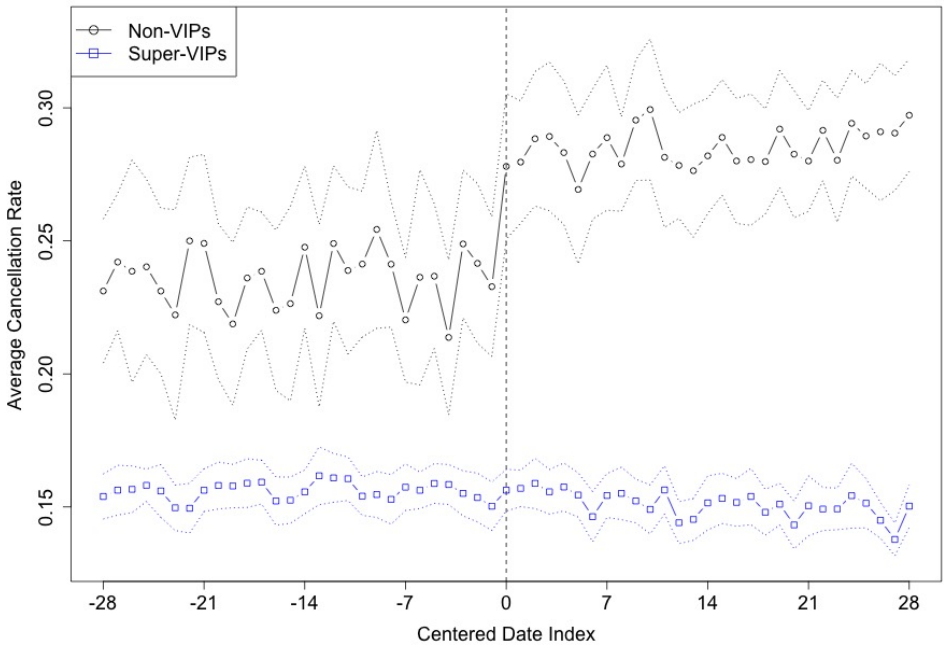
Figure 2.2 provides preliminary evidence of leakage. For the 33 treated cities, the average cancellation rates went up from 23.57% to 28.54%, or about 5%, for Non-VIPs who were charged commission after the new policy. There were no changes for Super-VIPs.

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 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.

[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 A.3.1 uses Beijing as an example to illustrate how the SCM estimates treatment effects

Figure 2.2. 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 statistical distribution of the SCM estimates are reported in Figure A.4 and Table A.2 in Appendix A.3.2.

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.

2.3. 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 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

[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.

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 2.3.1 for the variable construction). We will report the extent of disintermediation in Section 2.3.2. 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 2.3.3.

2.3.1. 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.

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

Drivers typically leave a trace in space and time as documented in Chapter 1. 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.

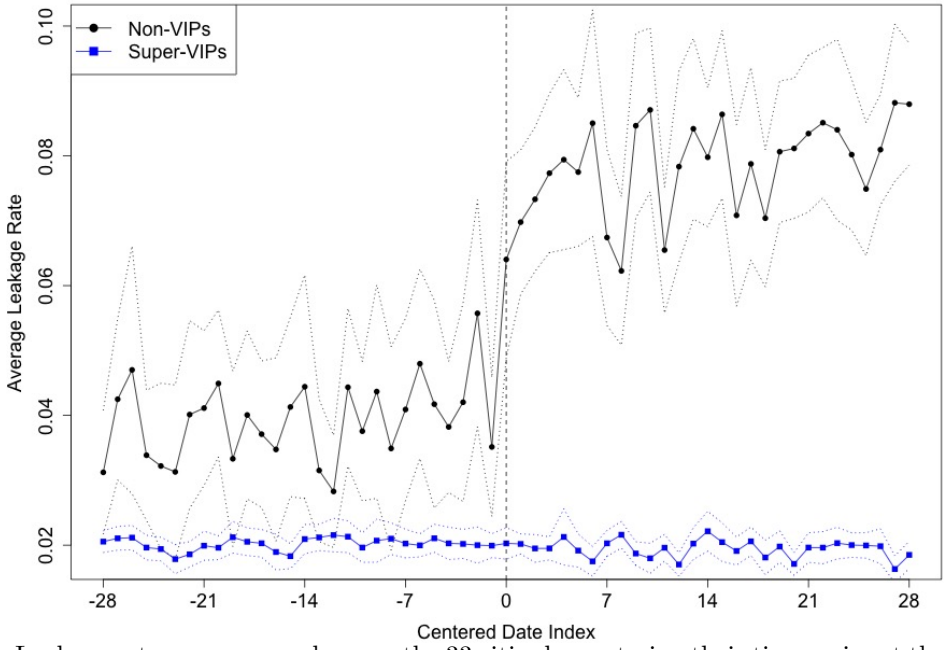
The GPS detection algorithm can achieve both high recall and precision (see Appendix A.2 for more details). Our approach is a more direct measure of disintermediation. It does not rely on complicated text analysis and other signals that are indirect measure. According to Xie et al. (2022), platforms would underestimate the scale of leakage if they only use keyword matching to create the measure of intention to disintermediate (Airbnb, 2022b; eBay, 2022a; 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 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.

[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.

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

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 2.3. 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.

2.3.2. 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. Figure 2.3 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 2.2 reports both the average leakage rates (i.e., the number of disintermediated transactions / the number of transactions) and cancellation rates (i.e., the number of cancellations / the number of transactions) across 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 2.2. 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.

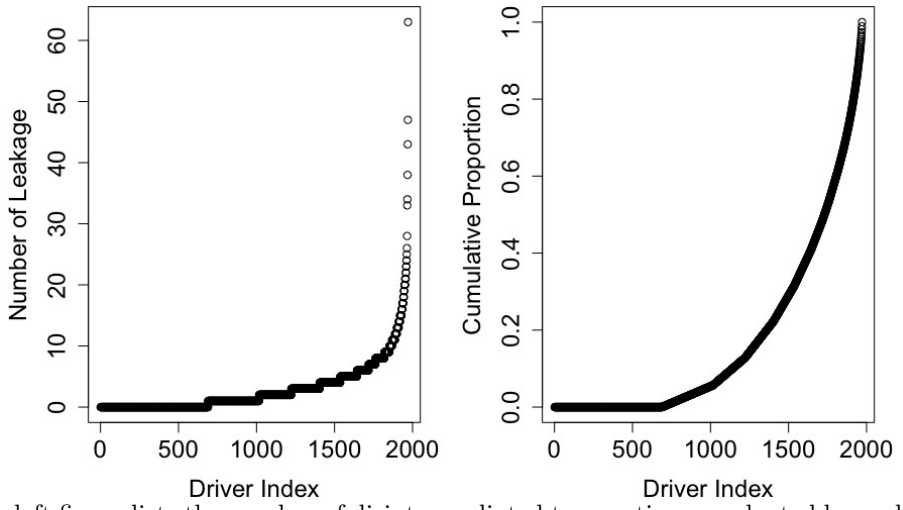
The simple pre-post analysis[22] at the city level gives us model-free evidence on the potential effect of commission launch on our direct measure of leakage. Such analysis assumes no structural changes other than the commission launch in these 33 treated cities. To alleviate the concerns on changes in time or changes in driver composition, we conduct two extra analyses: (1) Appendix A.3.1 documents a case study of Beijing and its neighboring cities using the city-level synthetic control methods (SCM); (2) Appendix A.4.1 uses the driver-level difference-in-difference (DiD) regressions. We discuss the pros and cons of these two approaches in Appendix A.4.2 regarding whether the additional analyses is suggestive or causal.

[22]Figure A.4 and Table A.2 report the statistical distribution of differences across cities (Appendix A.3.2).

2.3.3. 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.

Figure 2.4. 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.

The GPS detection algorithm identifies that 2/3 of drivers (see Figure 2.4) 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 (e.g., cash, escrow), item types (e.g., furniture), client types (e.g., enterprise users), distance, the time of service, the number of passengers, and the size of the vehicle requested.

[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.

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 (or 24.24% of transactions had commission for Non-VIP 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 2.5 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 2.7 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.

Figure 2.5. Leakage Rate by Price

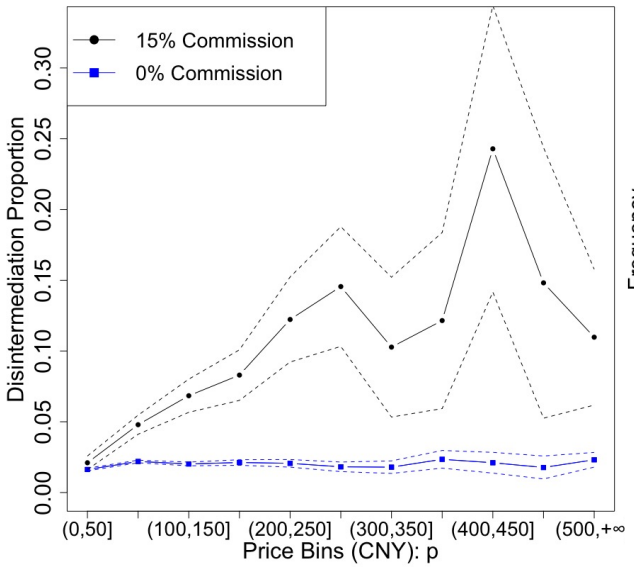


Figure 2.6. Histogram of Price

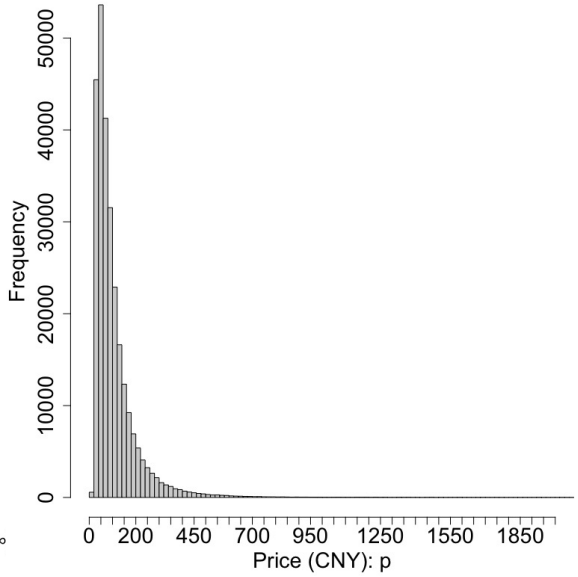


Figure 2.7. Leakage Rate by Coupon

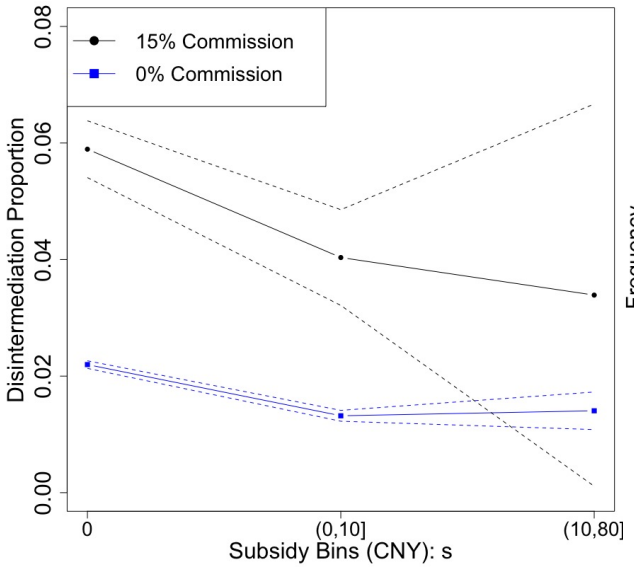
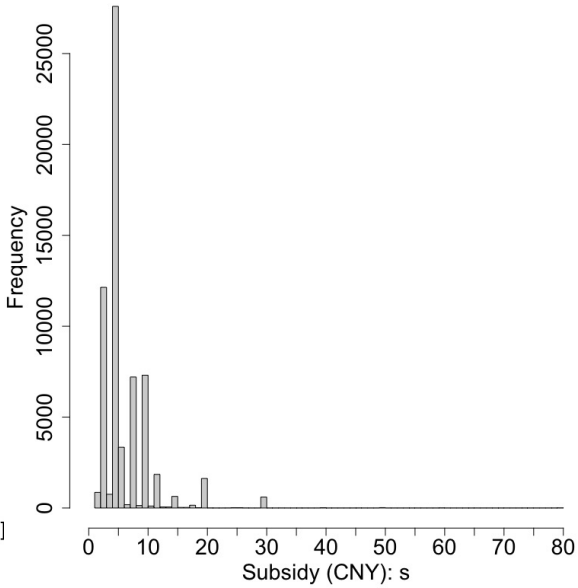


Figure 2.8. Histogram of Coupon



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.

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 2.3. 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%

Table 2.3 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 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 2.3.3 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 2.2.1, 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.

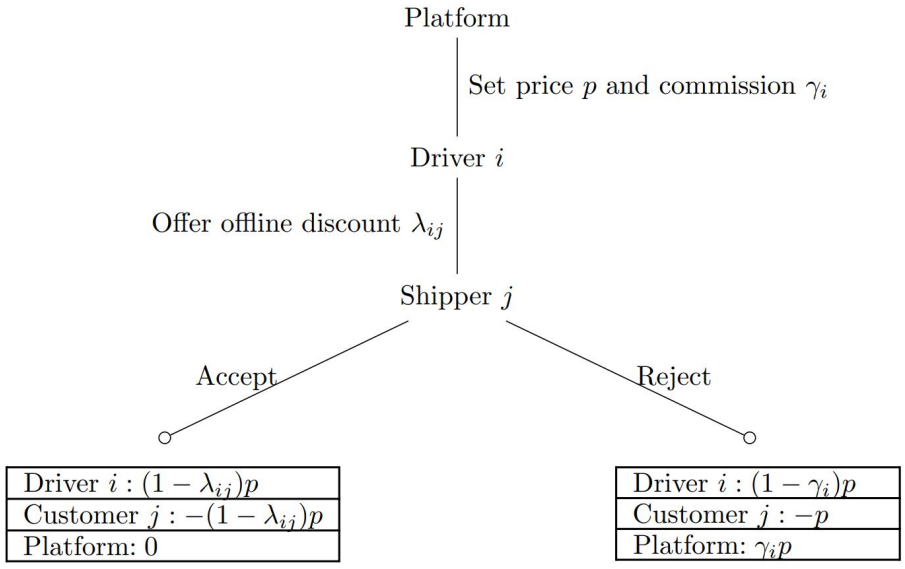
[24]The Road Traffic Safety of China prohibits the usage of cargo vans and trucks for rideshare service.

2.4. Economic Models

This section develops the model to investigate monetary incentives (Rochet and Tirole, 2006; Weyl, 2010), transaction costs (Coase, 1937; Williamson, 1987; Hagiu and Wright, 2023), 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 2.9 demonstrates the monetary transfers between the driver and customer when the transaction is conducted either on the platform or outside the platform:

Figure 2.9. 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. According to Spulber, 2019, platforms may set $\gamma_i p$ based on the elasticity of demand and supply, and transaction costs including moral hazard. We will formally set up the utility functions that take into account the price elasticities and transaction costs in Section 2.4.1.

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 A.1 in Appendix A.1). We also have anecdotal evidence of offline discounts in our on-demand cargo delivery platform.

2.4.1. 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:

$$\begin{aligned} \Pi_{i(t)}^0 &= \beta_i \cdot (1 - \gamma_{i(t)}) p_t \\ U_{j(t)}^0 &= u_{j(t)} - \beta_j \cdot (p_t - s_{j(t)}) \end{aligned} \tag{2.1}$$

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:

$$\begin{aligned} \Pi^1_{i(t)} &= \beta_i \cdot (1 - \lambda_{ij(t)}) p_t - h_{i(t)} \\ U^1_{j(t)} &= u_{j(t)} - \beta_j \cdot (1 - \lambda_{ij(t)}) p_t - h_{j(t)} \end{aligned} \tag{2.2}$$

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.

2.4.2. 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 $U^1_{j(t)}$.

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

[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.

(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.

2.4.2.1. Driver's Decision. The latent utility gain from leakage for driver i on trasaction t is

$$\begin{aligned}\Delta\pi_{ij(t)} &= \Pi^1_{i(t)} - \Pi^0_{i(t)} \\ &= \left[\beta_i \cdot (1-\lambda_{ij(t)})p_t - h_{i(t)}\right] - \left[\beta_i \cdot (1-\gamma_{i(t)})p_t\right] \\ &= \beta_i \cdot \underbrace{(\gamma_{i(t)} - \lambda_{ij(t)})p_t}_{\text{Offline Savings}} - h_{i(t)}\end{aligned} \tag{2.3}$$

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

$$L_{i(t)} = \begin{cases} 1 & \Delta\pi_{ij(t)} + \epsilon_{ij(t)} \geq 0 \\ 0 & \Delta\pi_{ij(t)} + \epsilon_{ij(t)} < 0 \end{cases} \tag{2.4}$$

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

2.4.2.2. Customer's Decision. Similarly, the latent utility gain from leakage for customers j on transaction t is

$$\begin{aligned}\Delta u_{ji(t)} &= U^1_{j(t)} - U^0_{j(t)} \\ &= [u_{j(t)} - \beta_j \cdot (1-\lambda_{ij(t)})p_t - h_{j(t)}] - [u_{j(t)} - \beta_j \cdot (p_t - s_{j(t)})] \\ &= \beta_j \cdot (\underbrace{\lambda_{ij(t)}p_t}_{\text{Transfer}} - s_{j(t)}) - h_{j(t)}\end{aligned} \tag{2.5}$$

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

$$L_{j(t)} = \begin{cases} 1 & \Delta u_{ji(t)} + \epsilon_{ji(t)} \geq 0 \\ 0 & \Delta u_{ji(t)} + \epsilon_{ji(t)} < 0 \end{cases} \tag{2.6}$$

The $\epsilon_{ji(t)}$ is the structural error that enters into the customer's decision of leakage. It include new information that are not observable to the econometrician.

2.4.3. 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 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:

$$Pr[L_t = 1] = Pr[\Delta\pi_{ij(t)} + \epsilon_{ij(t)} \geq 0, \Delta u_{ji(t)} + \epsilon_{ji(t)} \geq 0] \tag{2.7}$$

Bargaining (Nash, 1950; Rubinstein, 1982) happens when two agents can create a surplus together but requires a solution to split the surplus. We can approximate Equation (2.7) with Equation (2.8) by assuming that the driver-customer pair ij can find a $\lambda^*_{ij(t)}$[29] as long as the sum of utility gains from leakage is non-negative to reach the agreement of leakage (i.e., $Pr[\Delta\pi_{ij(t)} + \epsilon_{ij(t)} + \Delta u_{ji(t)} + \epsilon_{ji(t)} \geq 0]$). In other words, the driver and customer are assumed to be cooperative when they can create a net surplus together and divide it in a way to satisfy their minimum acceptable gains from an offline transaction.

[29]Appendix A.7 uses Nash Bargaining to microfound $\lambda_{ij(t)}$ to describe the the agreement process, including the special cases where one of the Individual Rationality (IR) constraints is binding when a $\lambda^*_{ij(t)}$ makes one party indifferent between transacting offline or online.

The $\lambda^*_{ij(t)}$ in this approximation[30] redistributes the joint surplus $(\Delta\pi_{ij(t)}+\Delta u_{ji(t)})$ to make it possible that both individual utility gains ($\Delta\pi^*_{ij(t)}$ and $\Delta u^*_{ji(t)}$) are non-negative before the idiosyncratic shocks kick in.

$$
\begin{aligned}
Pr[L_t = 1] &= Pr[\Delta\pi^*_{ij(t)} + \epsilon_{ij(t)} + \Delta u^*_{ji(t)} + \epsilon_{ji(t)} \geq 0] \\
&= Pr[\beta_i \cdot (\gamma_{i(t)} - \lambda^*_{ij(t)})p_t - h_{i(t)} + \epsilon_{ij(t)} \\
&\qquad + \beta_j \cdot (\lambda^*_{ij(t)}p_t - s_{j(t)}) - h_{j(t)} + \epsilon_{ji(t)} \geq 0] \\
&= F_\epsilon \left[\beta_i \cdot \gamma_{i(t)}p_t - (\beta_i - \beta_j)\lambda^*_{ij(t)} \cdot p_t - \beta_j \cdot s_{j(t)} - h_{i(t)} - h_{j(t)}\right]
\end{aligned}
\tag{2.8}
$$

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[31] based on real-time market or traffic conditions).

2.5. 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,

[30] An alternative approximation is to use the product of the individual leakage probabilities (i.e., $Pr[\Delta\pi_{ij(t)} + \epsilon_{ij(t)} \geq 0] \cdot Pr[\Delta u_{ji(t)} + \epsilon_{ji(t)} \geq 0]$), which requires the independence assumption.

[31] 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

we do not know 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[32](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 2.8 becomes Equation 2.9:

$$Pr[L_t = 1] = F_\epsilon \underbrace{\left[\beta_d \cdot \gamma_{i(t)} p_t - (\beta_d - \beta_s)\lambda^*_{ds} \cdot p_t - \beta_s \cdot s_{j(t)} - (h_{d(t)} + h_{s(t)})\right]}_{x'_{ij(t)}\theta} \tag{2.9}$$

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

2.5.1. Identification Strategies

Variation in the driver-side commission $\gamma_{i(t)} p_t$ (Figure 2.5) identifies β_d, the drivers' marginal utility for money. Variation in the customer-side subsidy $s_{j(t)}$ (Figure 2.7) identifies β_s, customers' marginal utility for money. Since we observe natural variation in the quoted price p_t (Figure 2.6), 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 2.3). 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

[32]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.

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[33] variation 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 $\theta = [\beta_d, \beta_s, \lambda^*_{ds(t)}, h_{d(t)} + h_{s(t)}]$

[33]Appendix A.5 provides some examples to illustrate the intuition for identification within an individual.

2.5.2. 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[34] 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 (2.9) 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:

$$\mathcal{L}(\theta) = \sum_{t=1}^{N} L_t \cdot ln \left[\frac{exp(x'_{ij(t)}\theta)}{1 + exp(x'_{ij(t)}\theta)} \right] + (1 - L_t) \cdot ln \left[\frac{1}{1 + exp(x'_{ij(t)}\theta)} \right] \tag{2.10}$$

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 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.

2.5.3. 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

[34] 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 A.5.

- 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 (2.9) can be estimated using the entire sample of transactions with

$$x_{ij(t)}^{'}\theta = -h - x_t^{'}\beta_{h_{(T)}} + \beta_d \cdot \gamma_{i(t)}p_t - \beta_s \cdot s_{j(t)} - (\beta_d - \beta_s)\bar{\lambda} \cdot p_t \tag{2.11}$$

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 2.4. 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[35]. 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 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 $\hat{\bar{\lambda}} = 0.074$.

[35]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 2.4. 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

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[36] 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

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

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 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.

2.5.3.1. 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 (2.9) with driver fixed-effects has

$$x_{ij(t)}^{'}\theta = -h_i - x_t^{'}\beta_{h_{(T)}} + \beta_d \cdot \gamma_{i(t)} p_t - \beta_s \cdot s_{j(t)} - (\beta_d - \beta_s)\bar{\lambda} \cdot p_t \tag{2.12}$$

Column (3) and Column (4) in Table 2.4 report the estimates. 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.

2.5.3.2. Customer Heterogeneity in Transaction Costs. Similarly to Section 2.5.3.1, 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 (2.9) with both driver and customer fixed-effects has

$$x'_{ij(t)}\theta = -h_i - h_j - x'_t\beta_{h_{(T)}} + \beta_d \cdot \gamma_{i(t)}p_t - \beta_s \cdot s_{j(t)} - (\beta_d - \beta_s)\bar{\lambda} \cdot p_t \tag{2.13}$$

Table A.4 in Appendix A.6 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.

2.5.4. 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 A.5 in Appendix A.6). 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 A.5 of Appendix A.6 reports the model estimates with heterogeneous price sensitivity. We use these rich estimates to conduct the counterfactual analysis in Appendix A.7 to guide the discussion of policy implications in Section 2.6.3. 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.

2.6. 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.

2.6.1. 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.

2.6.2. 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[37]. However, monitoring conversations faces many limitations in on-demand services. It not only triggers privacy concerns[38], 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[39]. 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.

[37]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.

[38]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).

[39]Table 2.5.3.1 shows that more passengers are associated with the lower hassle and higher leakage.

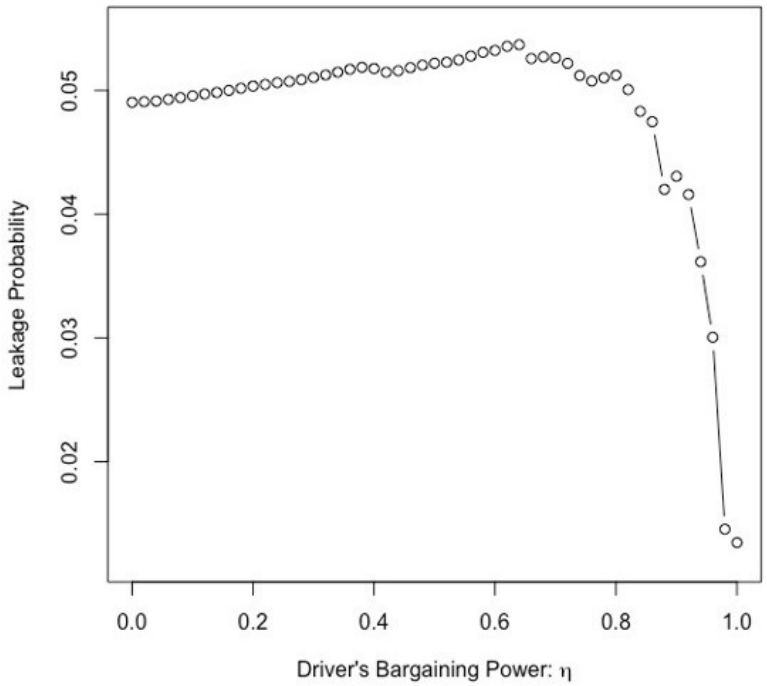
2.6.3. Matching Policy

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' hassle (see Section 2.5.3.1 for estimates). Given the historical information[40], 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 A.7 shows that the likelihood of leakage is higher when drivers have just slightly stronger bargaining power than customers (see Figure 2.10). 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 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., 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

[40]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.

Figure 2.10. Counterfactuals under 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 A.7 for the micro-founded model that uses Nash Bargaining as the solution concept).

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 A.7.3 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.

2.7. Conclusion

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 more 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 cannot get any job assignment for on-demand delivery without disclosing their locations. Moreover, no penalties were in place 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, 2022b).

CHAPTER 3

High-energy Ad Content: A Large-scale Investigation of TV Commercials (joint with Joonhyuk Yang, Lakshman Krishnamurthi, and Purushottam Papatla)

3.1. Introduction

Although consumers' responses to particular choices of ad content have long been studied (e.g., Olney, Holbrook, and Batra (1991); Preston (1982)), only recently have researchers started to use computational methods with large-scale, unstructured, nonlaboratory data. With increasing progress in data and methodology, marketers now have the opportunity to better understand the influence of ad content (which is inherently unstructured) on consumer behaviors, leading to better design of ads.

In this study, we focus on one aspect of ad content that has captured both researchers' and practitioners' attention, namely, high-energy stimuli in advertising. The energy of an ad is related to how stimulating it is. While there is no standard definition of "energy" in advertising, Puccinelli, Wilcox, and Grewal (2015) say: "High-energy commercials are television ads that are active, exciting and arousing for the viewer to experience" (p. 1). High-energy commercials are reported to have become prevalent in the TV advertising market. For instance, Puccinelli, Wilcox, and Grewal (2015) found that more than 80% of commercials on Hulu were rated as relatively energetic. Our investigation of Super Bowl

ads as well as TV commercials on the major broadcast networks also confirms that the energy level of ad content has increased.

Motivated by these observations, we examine whether the use of high energy in TV ads affects viewer behavior in terms of how long ads are tuned in. Studies in marketing have demonstrated that the inclusion of highly arousing stimuli increases viewers' attention to ads (e.g., Belanche, Flavián, and Pérez-Rueda (2017)). Other research, however, points to the opposite, namely that high energy can yield a negative outcome under certain conditions (e.g., Puccinelli, Wilcox, and Grewal (2015)). Given these different possibilities, we empirically evaluate whether and how higher energy in TV ads, overall, is associated with viewers' ad-tuning rates.

A key differentiating feature of our research is that, unlike previous studies conducted in lab settings, we utilize a large-scale data set of actual TV commercials tuned in at people's homes[1]. Specifically, our data set contains detailed information on the insertions of over 27,000 ads on the five major broadcast networks in the United States (ABC, CBS, FOX, NBC, and The CW) during the three-year period between September 2015 and August 2018. In all, our data set includes over a million insertions of the ads with information on the network, program, date, and time of each airing, as well as the video of each ad. For each insertion of an ad, we have data on the percentage of viewers who stayed tuned in for at least 25%, 50%, and 75% of the ad's duration. We use these measures to relate the energy in ad content to the tuning rate of ads. These measures are relevant to advertisers because so-called call-to-action or selling arguments typically come at the

[1]We distinguish the term "tuning" from "viewing," "watching," and "attention." Tuning in when an ad is played does not necessarily mean viewing, watching, or paying attention to the ad (e.g., see McGranaghan, Liaukonyte, and Wilbur (2022) on how the rates vary for these behaviors). In our data, we observe only tuning.

end of ads, following attention-grabbing content in the earlier parts (Teixeira, Picard, and El Kaliouby (2014)).

One empirical challenge is to measure each ad creative's energy level. Employing human coders would be both costly and infeasible given the large number of ads we aim to investigate, especially when many coders would be needed to rate each ad to ensure measurement reliability. Human ratings are also susceptible to individual factors and contextual factors such as the specific location, time of measurement, and pre-existing mood. We overcome this challenge using an algorithm-based approach. Specifically, we operationalize energy in ad content by adapting a widely used proprietary approach from the music streaming service Spotify. Spotify defines energy as "a perceptual measure of intensity and powerful activity released throughout the track. Typical energetic tracks feel fast, loud, and noisy." Notably, Spotify's energy measure has been formally investigated in computer science (Ferraro, Bogdanov, and Serra (2019); Ng and Mehrotra (2020)), sociology (Askin and Mauskapf, 2017), and marketing (Boughanmi and Ansari, 2021).

Nonetheless, the scalability and efficiency gains from using an algorithm-based approach also come with a cost. First and foremost, we need a better understanding of the relationship between the Spotify energy metric (an objective measure based on low-level auditory features) and viewers' perceived affect in our context. To this end, we conducted a separate analysis using data collected via Amazon Mechanical Turk (MTurk). Borrowing the circumplex model of affect from (Russell, 1980), we find that our algorithm-generated energy measure is positively correlated with the two dimensions of affect, namely arousal and valence.[2] Citing (Mehrabian and Russell, 1974), (Di Muro and Murray, 2012) define

[2]The arousal-valence framework from Russell (1980) provides a mapping of emotions onto a two-dimensional space, where the two axes are defined as arousal and valence. The model has been widely

arousal as "the subjective experience of energy mobilization, which can be conceptualized as an affective dimension ranging from sleepy to frantic excitement." Valence, on the other hand, is "the extent to which an affective state is positive or negative." We find that the correlation is much stronger with arousal (.46–.66) than with valence (.02–.21). Second, Spotify's algorithm was originally designed for music, so it can only measure the energy level of the auditory component of ads. Obviously, TV commercials consist of not only audio but video as well. Thus, if the energy level in an ad's audio differs substantially from that in its video or its overall energy level, our objective measure may not represent the actual consumer experience. However, results from the MTurk study confirm that perceived arousal from ad audio is well correlated with that from ad video (correlations ranging between .53 and .57) and, more importantly, also highly correlated with the overall perceived arousal from both ad audio and video (correlations ranging between .68 and .86).

Third, because of the proprietary nature of Spotify's measure of energy, we develop our own approach to predict the metric for the ads in our data set. We use a publicly available music data set of over 13,000 songs from the Free Music Archive (FMA), which includes the primitives of audio characteristics characterized by low-level attributes like frequency and spectral flux as well as the Spotify measure of energy for each song. We employ machine learning methods to relate the low-level auditory attributes of each song to the Spotify energy measure for that song. Next, we extract the audio from the videos of each of the ads in our data set and measure the primitive audio characteristics of the extracts. Finally, we use the relationship between energy and the primitive audio

adopted by researchers to measure affective state (for more discussion, see e.g., Fong, Kumar, and Sudhir (2021)).

characteristics derived from the FMA data set to estimate the audio energy in each ad creative.

We then empirically investigate the relationship between the tuning rates of ad insertions and their energy levels via a two-stage regression. The first stage includes fixed effects of the ads and insertion variables that include the network, program, date, time, and position of the insertion within the program. Controlling for these aspects of ad insertions is important because tuning rates are likely to be context dependent. The second stage takes the estimated fixed effects of the ads as proxies for their quality and regresses them on the derived Spotify energy measure and other creative characteristics of the ads, such as their mood and whether they include a song. The results indicate that, on average, more energetic commercials are more likely to be tuned in for longer or less likely to be avoided by viewers, supporting the positive association. However, the associations between energy levels in ad content and the extent of ad-tuning vary in both magnitude and direction across product categories and program genres.

To further assess the reliability of this evidence supporting the positive association of energy with tuning rates, we conduct a novel empirical investigation that draws its inspiration from ideal exogenous variation across or within ads in their energy levels, while holding other aspects constant. Specifically, we take advantage of the fact that people's preferences for energetic audio exhibit temporal variations over a day (Park et al., 2019). For instance, people generally prefer more energetic audio during the day and less energetic audio at night. Thus, showing an ad, which has a fixed energy level, at different times of the 24-hour day generates variation in the distance from its energy level to the level people generally prefer (i.e., the baseline) at those different times. We utilize

an auxiliary data set of Spotify listening sessions, the Music Streaming Sessions Dataset (MSSD; Brost, Mehrotra, and Jehan (2019)), to quantify this baseline. We then leverage the time-varying baseline levels to quantify the relative distance from the baseline to each ad's energy level, which generates within-ad variation in audio energy relative to the baseline. We relate the within-ad variation to tuning rates. This approach allows us to jointly estimate the ad fixed effects and energy effects in a single step. We find that higher energy is again associated with longer tuning in or lower ad avoidance, which provides additional support for the positive association.

Our research makes two contributions to the literature. First, we use large-scale data to empirically demonstrate that high-energy ad content is associated with higher tuning rate of ads. On this topic, previous research, heavily reliant on lab studies, finds a positive relationship between arousal and viewing time of ads (Belanche, Flavián, and Pérez-Rueda (2017);Olney, Holbrook, and Batra (1991)), while other studies find certain conditions in which increased arousal does not lead to an increase either in attention or in viewing time[3].

To the best of our knowledge, our work is one of a few studies that connect the tendency of ad avoidance to a particular aspect of ad content using large-scale, nonlaboratory data[4]; others include (Wilbur, 2008) and (McGranaghan, Liaukonyte, and Wilbur, 2022). More broadly, our study also adds to the rich marketing literature on ad avoidance in the context of television advertising (e.g., Danaher (1995); Deng and Mela (2018); Siddarth

[3]For instance, Puccinelli, Wilcox, and Grewal (2015) find that energetic advertising will lead to less watching when the context of the program is one of low arousal. Gorn, Tuan Pham, and Yatming Sin (2001) did not consider tuning rates but find that increased arousal may not always increase ad evaluation.

[4]In a detailed analysis of unstructured data in marketing, Balducci and Marinova (2018) describe unstructured data as nonnumeric, with no predefined representation of the construct of interest, as multifaceted, and as providing concurrent information.

and Chattopadhyay (1998); Teixeira, Wedel, and Pieters (2010); Tuchman, Nair, and Gardete (2018); Van Meurs (1998); Wilbur (2016)) by adding the role of a new dimension (i.e., energy in ad content) to the literature.

Second, we introduce a novel approach to measure the energy level of ad content, borrowing audio information retrieval methods from computer science and adapting the methods for a marketing context, thus adding to the literature on the use of unstructured, multimedia data in marketing research. The volume of marketing research that leverages unstructured multimedia data is ever increasing.[5] We believe that our study showcases the importance of parsimonious feature engineering of ad content in research relating it to consumer behavior. Algorithm-based approaches to ad content can easily end up with a complex, high-dimensional feature set. Given all the techniques available, one could extract thousands of auditory and visual elements of ads but find that very few are easily interpretable and capable of explaining meaningful variation in consumer behavior (e.g., McGranaghan, Liaukonyte, and Wilbur (2022)). In some sense, our approach goes in the opposite direction by carefully choosing a single human-interpretable variable, namely energy in ad content, examining the construct behind it, and investigating its relevance to consumer behavior. The work of Fong, Kumar, and Sudhir (2021) is also in a similar vein as they begin with a theory (i.e., Russell's arousal-valence model of affect) to guide their feature engineering.

[5]This research includes studies on text (e.g., Liu, Singh, and Srinivasan (2016); Netzer et al. (2012)), images (e.g., Jalali and Papatla (2016); Liu, Dzyabura, and Mizik (2020); Xiao and Ding (2014)), voice (e.g., Marinova, Singh, and Singh (2018); Xiao, Kim, and Ding (2013)), and videos (e.g., Li, Shi, and Wang (2019); Lu, Xiao, and Ding (2016)), among others.

3.2. Data and Ad Avoidance

We use data from multiple sources as summarized in Table 3.1. The primary data set for our research is provided by iSpot.tv (www.ispot.tv), which tracked over 9 million internet-connected televisions in the United States during the studied period of the three television seasons from September 2015 through August 2018.[6] The company relies on automatic content recognition (ACR) technology provided by other companies to detect whether an ad is being played on the screen and, if it is, which brand is being advertised and what creative is being shown. The ACR software installed on each smart TV tracks the ads either played live or time-shifted and played from a DVR-type recording within seven days of the original broadcast. Staff of iSpot.tv reported to us that about 6% of the viewership is time-shifted in our data. Because information is collected once every second from each tracked television, iSpot.tv is able to detect the exact instant at which the ad is interrupted. Specifically, the exact times of actions like fast-forwarding, switching to a different channel, pulling up the program menu, or turning the TV off while an ad is playing are recorded. The identity of the household member who takes those actions is unknown, as we do not have access to household-level data.

The data set we investigate in this research includes all nationally telecast ads available from iSpot.tv during the three television seasons from September 2015 through August 2018. The data set consists of over a hundred networks and covers all DMAs in the continental United States. Given the large number of ads over the three years, we limit our investigation to ads on five national broadcast networks: ABC, CBS, Fox, NBC, and

[6]Documentation from iSpot.tv describes that the population of tracked televisions is adjusted to ensure that the number of monitored televisions in each designated market area (DMA) and zip code of the country reflects the proportion of all TVs in the country present in that DMA and zip code.

Table 3.1. Data description

Source	Dataset	Observations	Obs. Period
iSpot.tv	Ad Insertion, Videos and Metadata	Ad insertion information (e.g., network, program, date, time), videos and metadata (e.g., duration, and other attributes)	Sep. 2015 - Aug. 2018
iSpot.tv	Ad-avoidance	Proprietary measure of the likelihood of ad interruption at each ad insertion	Sep. 2015 - Aug. 2018
adland.tv	Super Bowl Commercials	Videos of Super Bowl commercials	Super Bowl III (1969) - Super Bowl LIV (2020)
Free Music Archive (FMA)	FMA Database	Information on freely available songs (tracks) with their audio features	Songs (tracks) released in 1902 - 2017
Spotify	Music Streaming Sessions Dataset (MSSD)	Session-based log data on music choices, tracks and their audio features	July.2018 – Sep.2018

Notes: iSpot.tv: `https://www.ispot.tv`; adland.tv: `https://adlant.tv`; Free Music Archive: `https://freemusicarchive.org/`; Spotify's MSSD: `https://dl.acm.org/doi/fullHtml/10.1145/3308558.3313641`

The CW. Further, we consider only English-language programs and paid ads. Thus, ads by the networks promoting their own programming are not included.

Our unit of observation is a "creative insertion," where a "creative," or an "ad creative" or an "ad," is a particular ad execution for a brand, and an "insertion" refers to the unique network, program, date, and time combination during which an ad was inserted. The Geico brand, for example, has several ad creatives. A 30-second ad for this brand is thus a different creative from a 15-second ad for the same brand. The same creative could also have multiple insertions over different networks, different programs, different dates, and different times. Our data also include the actual video of each creative. In addition,

several variables coded by iSpot.tv are available. These variables include descriptors such as the emotion or emotions in the content, presence/absence of animals, and inclusion of popular music. Table 3.2 summarizes the metadata available for each ad.

Table 3.2. Ad-creative metadata

Variable	Description
Brand	Brand name associated with the ad-creative
Duration	15secs (baseline), 30secs, 60secs, 90secs, >90secs
Promotion	Indicator for whether the ad-creative includes a sales promotion message
Animal	Indicator for whether the ad-creative displays animals
Song	Indicator for whether the ad-creative has an accompanying popular song
Mood	Active (baseline): Indicator for whether the ad-creative is action-oriented; Emotional: Indicator for whether the ad-creative is emotional; Informational: Indicator for whether the ad-creative is informational; Funny: Indicator for whether the ad-creative is humorous; Sexy: Indicator for whether the ad-creative has a sexual theme

The data selected on the basis of our aforementioned criteria include more than 1 million insertions of over 27,000 ad creatives by 3,200 brands across 15 broad product categories[7]. Specifically, the data set records (1) the number of tracked TV sets tuned in for at least 3 seconds after the start of the ad ("start TVs") and (2) the number of TV sets still tuned in when at least 25%, 50%, or 75% of the ad has played ("end TVs"). Only a monitored TV that is counted among the start TVs is eligible to be counted among the end TVs. For example, when an ad creative that is 30 seconds long is aired, all monitored TVs that played the ad during the first 3 seconds would be the number of start TVs, and

[7]The 15 broad product categories coded by iSpot.tv are apparel, footwear, and accessories; business and legal; education; electronics and communication; food and beverage; health and beauty; home and real estate; insurance; life and entertainment; pharmaceutical and medical; politics, government, and organizations; restaurants; retail stores; travel; and vehicles.

the number that continued to play it for at least 75% of the length (23 seconds) would be the end TVs. Thus, the difference between the number of start TVs and the number of end TVs reflects the number of TV sets that interrupted the ad's showing before 25%, 50%, or 75% of its duration. Using the data set, we compute a measure of ad viewing called the ad-tuning rate as follows:

$$\text{Ad-Tuning Rate (ATR)} = \frac{\text{End-TV's}}{\text{Start-TV's}} \times 100. \tag{3.1}$$

Note that the "start TV" and "end TV" measures reset for every ad insertion. Suppose the first ad break of a program contains three ads. A television has to be tuned in to the ad for at least 3 seconds for the brand and ad creative to be identified. Thus, ads that are interrupted instantaneously are not included in the measurement. As an example, consider a 30-second ad. If 1 million televisions passed the 3-second threshold for the first ad in the ad break and 720,000 televisions are tracked at the 23-second mark, which is 75% of the length of the ad, then the ad-tuning rate at the 75% cutoff for this ad is 720,000/1,000,000=.72, or 72%. When the second ad is aired, the tracking starts afresh. Suppose this ad is also 30 seconds long, but this time 1.2 million televisions are tuned in for the first 3 seconds of this ad and 950,000 televisions are still tuned in at the 23-second mark. The ad-tuning rate for this ad is thus 950,000/1,200,000 or 79%. The second fact to note is that TVs that tune in to the ad after the first 3 seconds are not counted. Only televisions that are on for the first 3 seconds of the ad are tracked. Thus, the ad-tuning rate measure is specific to each ad and is unaffected by the number of TVs tuned in before or after the ad in question.

For the 75% cutoff, the mean ad-tuning rate across insertions of all the ad creatives in our data is 82.67%, and the standard deviation is 18.80%. At the 50% cutoff, the mean value is 94.33% with standard deviations of 10.48%, and at the 25% cutoff, the mean is 99.34% and the standard deviation is 3.53%. That is, the ad-skip rates in our data range between 5.67% and 17.33%, for cutoff values of 50% and 75%, respectively. These numbers may seem low, but they are not too far from those reported in previous studies. For instance, Danaher (1995) finds that television ratings drop by about 5% during commercial breaks in a data set from New Zealand. More recently, Tuchman, Nair, and Gardete (2018) report an ad-skip rate of about 5% in a data set from a Western European country. Using a TiVo data set from the United States, Deng and Mela (2018) report an ad-skip rate of about 15% for live viewing. Nonetheless, we provide some reasons for why the tuning rates in our study are higher than commonly held priors that suggest lower ad watching. First, those who immediately switch out within 3 seconds are excluded from the base because the ACR software needs this time to identify the ad creative. In particular, this exclusion partly explains why the view rate exceeds 99% with the 25% cutoff. For instance, the 25% duration of a 15-second ad is 3.75 seconds. Second, our measure omits a portion of time-shifted views using a DVR, for which ad skipping is more frequently observed (e.g., Deng and Mela (2018)). Third, the measure does not account for the population that watches TV programs via streaming. Because we use the same measure for all the ads that we investigate, however, these reasons will not distort our findings.

We use three additional data sets for our investigation:

(1) Free Music Archive: The FMA is a publicly available data set used widely in the music information retrieval (MIR) literature (Defferrard, Benzi, Vandergheynst, and Bresson, 2017). We use these data to develop an approach to reverse engineer Spotify's proprietary audio feature measurement methodology.

(2) Super Bowl: We collected the videos of all 3,077 ad creatives that aired during the Super Bowl between 1969 and 2020 from the website adland.tv. We use these data to explore the changes in the energy level of these commercials over a long period.

(3) Music Streaming Sessions Dataset: As of 2019, the MSSD is the largest publicly available data set for researchers to track consumers' preferences in streaming music and includes details of the musical tracks played and the specific hour and minute of the day at which the tracks are played (Brost, Mehrotra, and Jehan, 2019). The data set includes a log of a total of 3.7 million tracks that were played in 150 million listening sessions on Spotify, and we use a 20% random sample for computational tractability. We use these data to characterize the temporal patterns in the energy levels of streamed audio tracks.

3.3. Measuring Energy in Ad Content

The empirical question of interest is whether the association between the level of energy in television ads and the extent of ad tuning is positive or negative. To examine this association, we measure the energy level of each ad in our data set. In this section, we explain our operational measure of energy in ads using the Spotify audio energy measure. We then present data patterns from several descriptive analyses.

3.3.1. An Operational Measure of Energy in Ads

To measure the audio energy in TV commercials, we draw on work done by a company called Echo Nest, which was founded in 2005 as a research spin-off from the MIT Media Lab and acquired by Spotify in 2014. Echo Nest developed a proprietary approach to measure multiple characteristics of audio tracks (hereafter Echo Nest or Spotify attributes), which some consider "the current gold standard in MIR" (Askin and Mauskapf, 2017). The Echo Nest attributes include seven subjective measures, which are labeled acousticness, danceability, energy, liveness, instrumentalness, speechiness, and valence, and one objective measure, namely, tempo. Developers and researchers can use the Spotify application programming interface (API) to get the Echo Nest attributes of tracks that are available in Spotify's music database. However, we cannot use Spotify's API to estimate the audio energy in the TV ads because TV ads are not listed in the Spotify database.[8] Moreover, our focus is on the overall auditory energy of ads, where the audio includes not only background music but also other sounds (e.g., speech and nonmusical sounds). Therefore, we develop an alternative approach that relies on an open-source algorithm called Librosa (McFee, Raffel, Liang, Ellis, McVicar, Battenberg, and Nieto, 2015) and the FMA data set mentioned previously.

Librosa can be used to decompose audio into a large number (¿500) of low-level spectral and rhythmic audio primitives or features. These features include chroma, mel-frequency cepstral coefficients, root-mean-square energy, spectral bandwidth, spectral contrast, tonnetz, zero crossing rate, and others, for which statistics such as minimum, maximum,

[8]In our data set, 85% of the TV ads do not use music from tracks that are available on Spotify, and the other 15% contain only snippets of the songs, which precludes the direct use of the Spotify API even if our focus is solely on background music. The use of recorded music in TV ads involves paying copyright fees, which is the reason that very few TV ads use such music.

median, mean, standard deviation, kurtosis, and skew are reported. The FMA data set includes 13,129 tracks for which both the Librosa audio features and Echo Nest audio attributes are available.

Using the data set, we reverse engineer the Echo Nest attributes by building a model that predicts the Echo Nest attributes based on the Librosa features. We communicated with a developer of the Echo Nest attributes, asking if it would be appropriate to derive Echo Nest attributes from the Librosa features. The developer responded that (1) a prediction model using Librosa features could work in principle, and (2) it would be an approximation of the original model. We summarize the process below (see Web Appendix A for more details):

(1) Using the FMA data set, we employed machine learning methods to derive the relationship between the Echo Nest energy attribute and the Librosa features. We explored different model forms (e.g., nonlinear) and machine learning methods (e.g., deep learning) to choose the model that achieved the best out-of-sample prediction.

(2) We extracted the audio component of each ad creative in our iSpot.tv data using FFmpeg, which is an open-source project that can extract the audio from all types of video files, such as TV ads.

(3) We used Librosa to decompose the audio files from Step 2 into the 518 low-level spectral and rhythmic features.

(4) We used the trained machine learning models (from Step 1) to estimate the energy level of each ad creative using the Librosa features from Step 3.

By doing so, we implicitly assume that the relationship between the Librosa features and the Echo Nest attributes from the FMA data set holds for audio in general. We do this not only for energy but also for the other Echo Nest attributes. While the prediction precision varies across attributes, we are confident in reverse engineering the energy attribute. For instance, the out-of-sample R-square value for energy is about.8, and the root mean square error for this attribute is the lowest among all the attributes. Table 3.3 presents sample TV ads with the highest and lowest predicted audio energy levels (see Web Appendix A.3 for URL links to the commercials). In our own evaluation, the two sets of ads are substantially different in terms of how energetic we perceived the commercials to be, with some exceptions (e.g., the Colgate TV commercial is not particularly energetic). Of course, our evaluation is subjective. To better understand what our measure of energy represents, we need a more systematic approach, which we discuss in the next section.

Table 3.3. Example of TV commercials with highest and lowest energy levels

	Brand (Commercial)	Predicted value	URL
	Highest energy		
1	Omega (Spectre: Revealing the 007 Watch)	0.8523	Link
2	Nike (Breaking 2)	0.8396	Link
3	Fage Yogurt (So Rich)	0.8367	Link
	Lowest energy		
1	Ralph Lauren Fragrances (Love)	−0.0211	Link
2	Center for Biological Diversity (Polar Bear)	−0.0089	Link
3	Clorox (Bleach It Away: Distance)	−0.0087	Link

Note: Table lists the examples of TV commercials with the highest and lowest values of energy in our dataset.

3.3.2. Data Patterns

We present results from three exploratory data analyses that describe the variation of the auditory energy level in ads across time, product categories, and program genres. These analyses document the following patterns using our data sets: (1) The energy level of TV ads is growing year after year with distinct intraday and intraweek temporal patterns. (2) Energy levels are heterogeneous across product categories. (3) Energy levels vary substantially across ads placed in different genres of TV programs. These findings are discussed in this section.

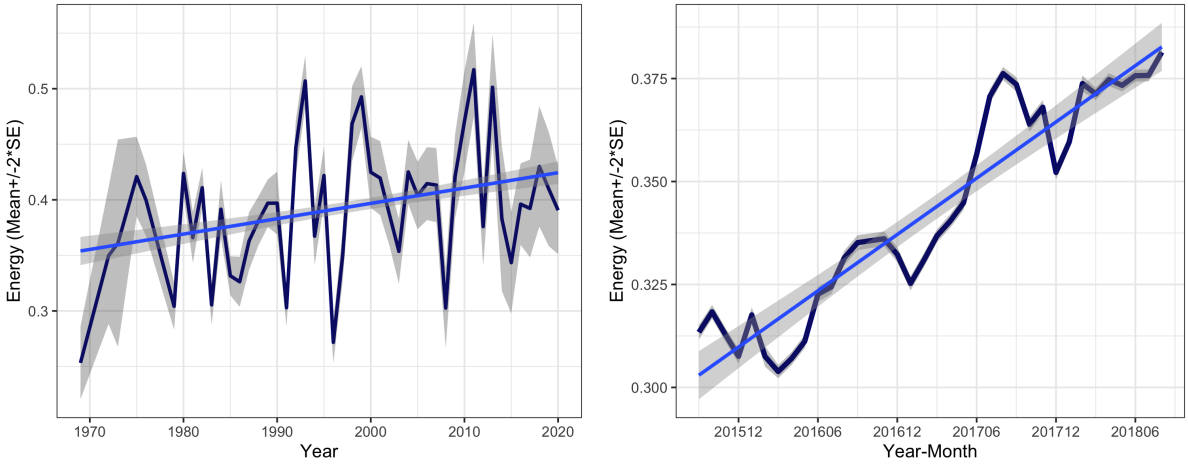
3.3.2.1. Energy levels over time. Figure 3.1, Panel A, depicts the average monthly energy levels of all ad creatives from the iSpot.tv data set, showing that the energy levels are increasing over these three years. The linear fitted value of the energy level is.278 in September 2015 and.367 in August 2018, which indicates a 33% increase during that period. To further investigate the temporal pattern over a longer horizon, we extracted the audio component from Super Bowl ads spanning six decades, and we plot their energy levels in Figure 1, Panel B. Each year includes ads with high energy levels and ads with low energy levels, as seen in the dispersion, but there is a clear increasing trend in energy levels over the years. The linear fitted value of the energy level in 1969 is.354, and that in 2020 is.424, which indicates a 20% increase during the time period. One caveat is that Super Bowl ads are unique because advertisers tend to employ creative executions that are more dramatic than usual, as seen in the higher energy levels for Super Bowl ads.

Next, Figure 3.2 illustrates the change of energy levels in the TV ads by hours within a week using the iSpot.tv data set. The temporal patterns are distinct: the energy level dips around 5 A.M. and peaks around 11 P.M. The level of some ads rallies around 6 A.M.

on weekdays, but the phenomenon does not hold on weekends. In general, commercials on weekend afternoons (12 P.M.–6 P.M.) are more energetic than those aired on weekdays.

Figure 3.1. Trends in energy levels of TV commercials

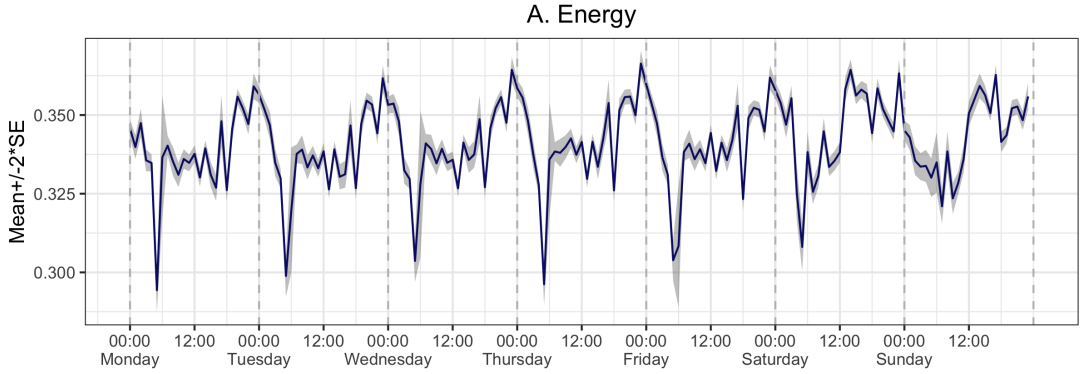
(A) Super Bowl ads: 1969-2020 (B) Major broadcast networks: 2015-2018



Notes: The two figures show that the average energy-level of TV ads is growing year over year: (a) plot on the left demonstrates the average auditory energy level of the Super Bowl ads from Super Bowl III (1969) to Super Bowl LIV (2020); (b) plot on the right illustrates the average auditory energy level of the ads shown at the five major broadcast television networks (ABC, CBS, FOX, NBC, The CW) in the iSpot.tv dataset (Sep. 2015 - Aug. 2018).

3.3.2.2. Energy levels across product categories. As shown in Figure 3.3, Panel A, energy levels in ads vary across the 15 product categories in the iSpot.tv data set. The product categories are ranked on the basis of their mean energy levels. The top three are vehicles (mean =.371), life and entertainment (mean =.367), and retail stores (mean =.365), while business and legal (mean =.283), politics, government, and organizations (mean =.258), and education (mean =.229) make up the bottom three. Typically, ads for trucks, sport-utility vehicles, entertainment, and retailers tend to be fast, loud, and

Figure 3.2. Temporal patterns in the energy levels of TV commercials

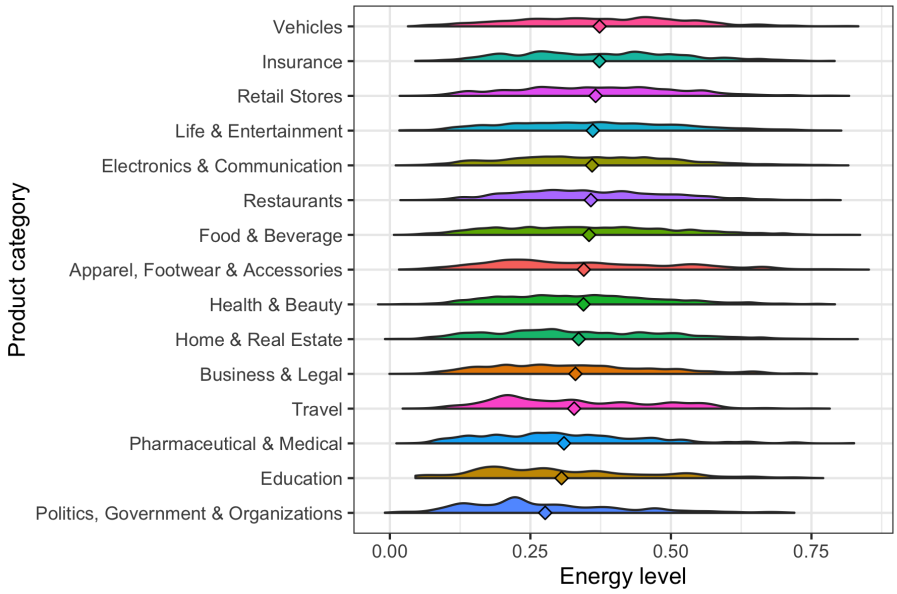


Notes: The figure displays the temporal patterns in the auditory energy level of the ads shown at the five major broadcast television networks (ABC, CBS, FOX, NBC, The CW) in the iSpot.tv dataset (Sep. 2015 - Aug. 2018). The solid line represents the average energy level of all ad-creatives for each hour of the day.

noisy. In contrast, the bottom three categories, which typically use somber and informational creatives, have low energy. Specifically, ads for business and legal services as well as those for educational institutions and governments are low on energy. The mean differences between the top and the bottom three categories are statistically significant ($t = 156.56, p < .001$).

3.3.2.3. Energy levels across program genres. The three main program genres are sports, entertainment, and news and information. Two other genres in the data set, labeled "various programs" and "infomercial," are dropped in this and subsequent analysis because their sample sizes are negligible. Figure 3.4, Panel B, displays the distributions of energy levels for ads placed in these three genres. Ads aired during sports programs have the highest energy on average (mean = .355), followed by ads aired during entertainment programs (mean = .318). In contrast, the genre with the lowest energy on average are ads placed in news and information programs (mean = .304). The mean differences are

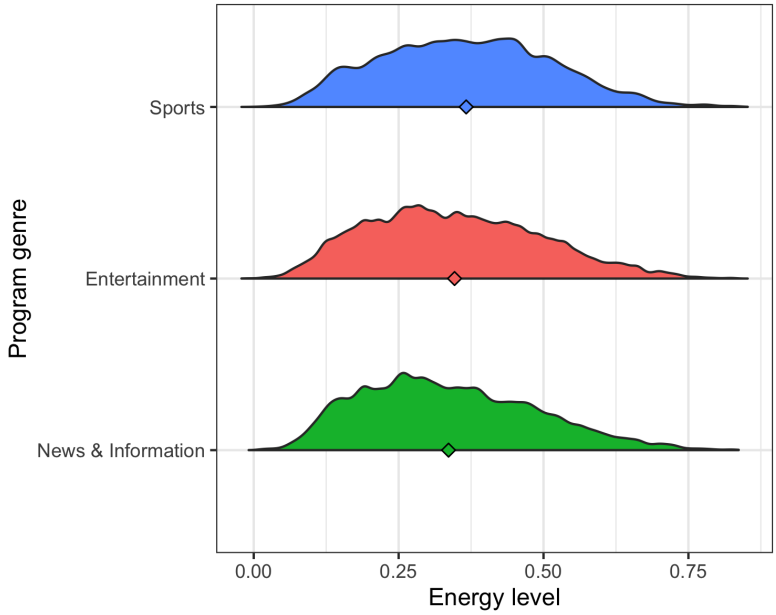
Figure 3.3. Energy levels of TV commercials by product categories



Notes: The figure plots the distributions of energy level for each product category. The product categories are ranked based on their mean energy levels. Commercials for vehicles, insurance and retail stores are most energetic on average, while those for politics, education and pharmaceutical have relatively lower energy levels.

statistically significant for sports versus entertainment ($t = 70.79, p < .001$) and news and information ($t = 90.84, p < .001$). Also, entertainment is significantly different from news and information ($t = 45.38, p < .001$). We expect sports programs to be more competitive and intense, whereas news broadcast programs are less upbeat, and entertainment programs will fall somewhere in between. The fact that the energy levels of TV ads in our data set vary in the same order across program genres indicates that advertisers may match their ads with the energy level of TV programs.

Figure 3.4. Energy levels of TV commercials by program genres



Note: The figure displays the distributions of energy level for ads placed in each program genre.

3.4. Algorithm-Generated Versus Human-Perceived Energy

What does Spotify ad audio energy represent? To answer the question, we borrow the arousal-valence framework developed by Russell (1980). The model provides a mapping of emotions onto a two-dimensional space, where the two axes are defined as arousal and valence that share the same aforementioned definitions from Di Muro and Murray (2012). For instance, excited, delighted, and happy emotions are mapped into the first quadrant as they are high on arousal and the valence is positive. Alarmed, afraid, and angry are in the second quadrant, which is high on arousal and has negative valence. On the other hand, sad, tired, and bored fall into the third quadrant as they are low on arousal and the valence is negative. Finally, emotions such as calm, relaxed, and contended are mapped in the fourth quadrant, being low on arousal with positive valence.

Our goal is to understand how our algorithm-based, objective measure of energy in ad content is related to the subjective measures of affect along the arousal and valence dimensions. Specifically, we compare human-perceived energy in ads evaluated by MTurk participants to the Spotify audio energy estimated for these ads on the two dimensions. We report a full description of our MTurk study in Web Appendix B. Here, we briefly explain the measurement and the results.

The measurement scales used in the MTurk study were borrowed from Puccinelli, Wilcox, and Grewal (2015) and Belanche, Flavián, and Pérez-Rueda (2017). We used multiple items with seven-point bipolar scales to assess arousal and valence. Items 1–7 ("Not energetic" to "Energetic," "Dull" to "Exciting," "Not animated" to "Animated," "Inactive" to "Active," "Relaxed" to "Stimulated," "Calm" to "Excited," and "Unaroused" to "Aroused") were used to load on the arousal factor, while items 8–12 ("Unhappy" to "Happy," "Displeasure" to "Pleasure," "Feel bad" to "Feel good," "Sadness" to "Joy," and "Negative" to "Positive") were used to load on the valence factor. We reverse coded some of the items to prevent straight-lining and randomized the order of the items to guard against acquiescence bias. The internal consistencies (Cronbach's alpha) were all above.95 for both arousal and valence.

We collected data from 2,342 MTurk participants on a randomly selected sample of 138 ad creatives[9]. Participants were randomly assigned to one of three conditions: (1) they can only hear the audio of ads without video (audio-only), (2) they can only watch the video of ads without audio (video-only), or (3) they can hear and see both audio and video

[9]We randomly selected 30 ad creatives for a pilot study and 110 ad creatives for the main study. Two of the 30 ads in the pilot study were excluded because of a copyright issue regarding the music that accompanied the ads. The cost was about $2,500 to rate the energy level of 138 ads through MTurk, so evaluating all 27,000 ads would have cost at least half a million USD.

of ads (audio+video). By doing so, we aim to understand how similarly or differently the MTurk participants perceive the levels of arousal and valence from different stimuli in ad content (i.e., auditory, visual, or both).

Overall, we find that the Spotify audio energy is related to the level of arousal in ad content. The correlations between the Spotify audio energy and human-perceived arousal in ad content from the audio-only, video-only, and audio+video conditions are.44,.38 and.46, respectively, in the main study (Figure W2 in the Web Appendix). We cannot say whether these correlations are high, but importantly, Spotify audio energy is consistently correlated with multiple measures of arousal (e.g., energetic vs. non-energetic; exciting vs. dull; active vs. inactive; stimulated vs. relaxed; aroused vs. unaroused; see Figures W4 to W6 in the Web Appendix). In contrast, the correlation between Spotify audio energy and valence from the audio-only, video-only, and audio+video conditions are lower at.14,.21, and.13, respectively, in the main study (Figure W3 in the Web Appendix). This finding is reassuring as the correlation between energy and valence derived using our machine learning algorithm from the FMA data set is.22 (Table W6 in the Web Appendix).

In addition, the correlations between the levels of perceived arousal from the three conditions are.57 between audio-only and video-only,.74 between audio-only and audio+video, and.86 between video-only and audio+video. We believe this high level of correlation is due to close alignment between the arousal stimuli in the audio and video in ads. The high correlation in the levels of arousal from the audio-only and audio+video conditions suggests that the Spotify energy measure can capture the overall arousal level and not just that from the audio in ad content.

We also obtain additional insights on the human-perceived arousal in ads. For instance, we find that Spotify audio energy is significantly more correlated with the arousal in background sound than with the arousal in human speech in the audio-only condition (.53 vs. .33; z = 2.616; p =.009) and the audio+video condition (.48 vs. .35; z = 1.691; p = .091; see Figure W7 in the Web Appendix). This finding is not particularly surprising, because Echo Nest originally designed the algorithm to summarize the attributes of music rather than speech.

Additional reinforcing evidence of the positive correlation between energy and arousal comes from the keywords provided by survey participants when they were asked to "list some words that come to mind when you think about energetic ads" (see Web Appendix B.1.5). The top five keywords mentioned by participants in the audio+video condition were "fast," "music," "movement," "upbeat," and "exciting." In the audio-only condition, participants listed "music," "upbeat," "fast," "loud," and "exciting," whereas in the video-only condition, they listed "movement," "fun," "active," "fast," and "exciting" (Figure W9 in the Web Appendix). The emotions represented by the keywords appear to fall in the high arousal–high valence quadrant. Note that the keywords from the audio+video condition appear to be the union of the keywords in the audio-only and video-only conditions. The results clearly suggest that both auditory and visual elements contribute to the overall energy level in ads, as measured by arousal. In the Web Appendices B.2 and B.3, we further investigate auditory and visual correlates of energy in ad content.

3.5. Associations Between Ad Energy and Ad-Tuning Rate

In this section, we explore the association between the ad-tuning rate and the energy level of ad content. The dependent variable of interest is the ad-tuning rate, ATR_{it}, measured as the percentage of TVs tuned in to an ad for at least 75% of its duration (we also discuss the results from the 50% and 25% cutoff values). The index i represents a unique creative execution of an ad, and the index t represents the insertion, which is a unique combination of network, program, date, and time. The independent variable of interest is Energy_i, which is the energy level of an ad creative computed using the first 75% of its duration. This variable accounts for the fact that for a tuning rate measured for a 75% cutoff, viewers would have been exposed to the ad's energy during the first 75% of its duration.[10] To recover the association between ATR_{it} and the energy level of ad creative i, Energy_i, we employ two empirical strategies: a between-estimator approach and a within-estimator approach.

3.5.1. A Between-estimator Approach

Here we investigate the association between ATR_{it} and Energy_i using a two-stage model. In the first stage, we use a fixed-effects model to quantify the ad-creative fixed effects (denoted by δ_i), which we interpret as an insertion-invariant measure of the "quality" of ad creative i in terms of ad tuning. Specifically, the ad-creative fixed effects capture all the components that constitute an ad creative and explain audiences' ad-tuning behavior. In the second stage, we project various aspects of ad creative i, including Energy_i, to the

[10]The correlation between the energy levels measured using the first 75% and 100% of ads' duration is about .962. The results are qualitatively unchanged when 100% of ad duration is used.

estimated δ_i from the first stage. Next, we explain the specification and the results of each stage.

First stage: Measuring ad-creative fixed effects. We first fit the model below:

$$\text{ATR}_{it} = \delta_i + X_t'\gamma + e_{it}. \tag{3.2}$$

where δ_i represents the ad-creative fixed effects, X_t is a set of variables that summarize the insertion of i, and e_{it} is an error term. For X_t, we include network fixed effects, program fixed effects, year-week fixed effects, day-of-week fixed effects, day part fixed effects[11], and position fixed effects. Network fixed effects are for the five broadcast networks, while program fixed effects are specific to each program, such as The Big Bang Theory (CBS) or This Is Us (NBC). Multiple networks carry some events like presidential debates and political conventions, and they also occasionally rebroadcast other networks' programs, which allows us to identify both network and program fixed effects. "Position" refers to the pod and pod position in which the ad is shown. A program consists of several ad breaks called pods, and each pod has several ads represented by their position in the pod, for example, the first ad in the first pod, the third ad in the second pod, and so on. Previous studies have demonstrated that viewing behavior can systematically differ across the positions of ads within a program (e.g., Danaher (1995); Van Meurs (1998)). We include a fixed effect for each pod and for each position within each pod, collapsing pods after the tenth position at ten.

[11]A 24-hour day is divided into nine day parts on television: early morning (Monday–Friday 6–10 A.M.), daytime (Monday–Friday 10 A.M.–4 P.M.), early fringe (Monday–Friday 4–8 P.M.), prime time (Monday–Saturday 8–11 P.M., Sunday 7–11 P.M.), late fringe P.M. (Monday–Sunday 11 P.M.–12 A.M.), late fringe A.M. (Monday–Sunday 12–2 A.M.), overnight (Monday–Sunday 2–6 A.M.), weekend day (Saturday–Sunday 6 A.M.–1 P.M.), and weekend afternoon (Saturday 1–8 P.M., Sunday 1–7 P.M.).

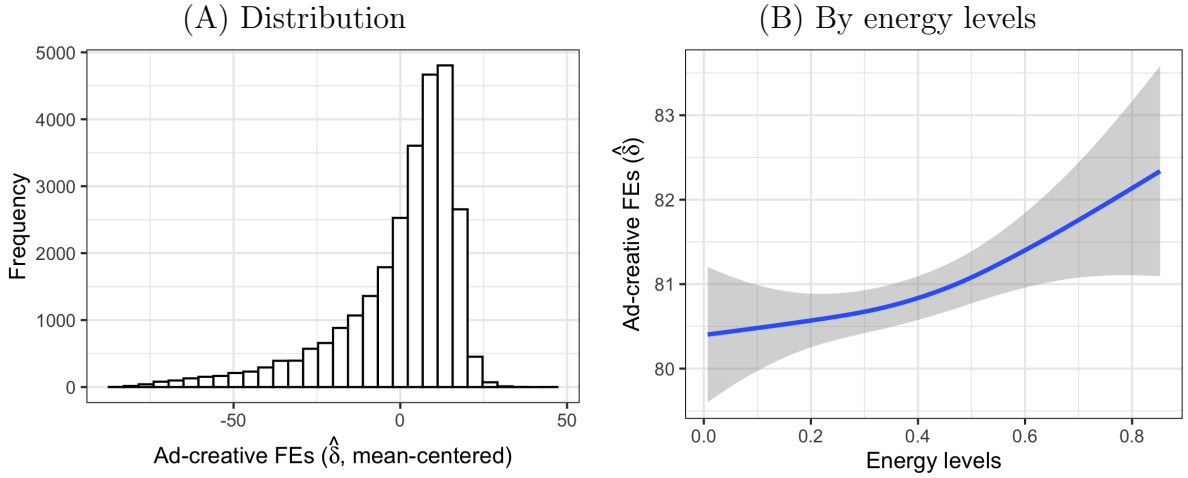
To assess the relative ability of various fixed effects to explain ATR_{it}, we compute the variation in ATR_{it} explained by our fixed effects when used either alone or in a combination. To ensure reliable estimation of the ad-creative fixed effects, for this part of the study, we drop about 8,200 ad creatives with fewer than five insertions in total during the three-year period. These ads account only for 1.6% of insertions in our data.[12] We find that ad-creative fixed effects alone explain 55% of the variation, which is far greater than that explained by network fixed effects (.17%); network and program fixed effects (2.20%); year-week, day of week, and day part fixed effects (2.59%); and network, program, year-week, day of week, and day part fixed effects (4.45%). It is logical that the strongest explanation of ATR_{it} is provided by the ad-creative fixed effects. When all the fixed effects are used, we explain about 57% of the variation in ATR_{it}. We report this variation for all possible combinations of fixed effects in the Web Appendix (Tables W11 and W12).

We use weighted least squares with the square root of the number of viewers as weights to estimate the full regression in Equation 3.2 (i.e., the variance of the observed ad-tuning rate is assumed to be inversely proportional to the number of viewers). The use of these weights accounts for the relative precision of ad-tuning rates that are measured from different numbers of viewers. The specification of the regression thus provides estimates of the ad-creative fixed effects while controlling for the network and program on which the ad is shown, and time and ad position. Higher δ_i indicates that ad creative i is intrinsically more likely to be tuned in to. Figure 3.5, Panel A, presents the distribution of the estimated δ_i values and shows that there is substantial variation in the estimated

[12]The results are robust when we drop ad creatives with fewer than 10, 20, or 30 insertions in total, although the reduced sample size affects the precision of estimates.

ad-creative fixed effects, which we exploit in the next stage. The estimates of other fixed effects are reported in Web Appendix C.

Figure 3.5. Estimated ad-creative fixed effects



Notes: The left panel reports the distribution of ad-fixed effects from the first stage regression. Higher $\hat{\delta}_i$ indicates that ad-creative i is intrinsically less likely to get avoided. The right panel reports the association between estimated ad-creative fixed effects and energy-level. The blue line reports a smooth spline and the grey band displays confidence interval.

Second stage: Decomposition of ad-creative fixed effects. As explained, the estimated ad-creative fixed effects capture all the components that constitute an ad creative (e.g., brand, the message, and ad duration) *and* explain audiences' ad-tuning behavior. In Figure 3.5 Panel B, we plot the distribution of the estimated ad-creative fixed effects and the corresponding estimated energy levels of the ads and find a positive relationship. We next formally investigate whether there is still a meaningful association between the energy level of ads and ad-tuning rates even after controlling for other factors of the ad. Specifically, we utilize the variables in Table 2 that measure the characteristics of the content of ad creatives and run the following regression:

$$\hat{\delta}_i = \alpha + \beta \cdot \text{Energy}_i + Z_i'\eta + \epsilon_i, \tag{3.3}$$

where Z_i includes all the variables shown in Table 3.2.

The dependent variable in Equation 3.3 is measured with error, very likely heteroskedastic, because it is an estimate from the first-stage regression. Following Lewis and Linzer (2005), we use a feasible generalized least squares (FGLS) method to account for the errors in the dependent variable and the model error in Equation 3.3.[13] We provide more details on this procedure and the estimation results from other approaches in Web Appendix D.

Table 3.5 reports the estimation results of Equation 3.3. The "brand" in Table 2, which is the brand that was being advertised in the ad creative, is modeled using brand fixed effects. It is clear from the improvement in fit in Table 3.5 that the model would be misspecified without controlling for the brand. We find that the energy level is positively associated with the ad-creative fixed effects, and the association is statistically significant when we include brand fixed effects. Thus, ad creatives with higher energy levels will have larger fixed effects in Equation 3.2, which, in turn, means that they will have higher tuning rates as well. [14]

[13] An alternative approach is to use a hierarchical model in which we combine Equations 3.2 and 3.3 into a single unified model with a random effect, rather than having fixed effects. This approach could result in a more accurate estimation of the standard error for energy.

[14] To account for a potential nonlinear relationship between energy and ad-creative fixed effects (Figure 4, Panel B), we also use alternative functional forms for the energy variable: a log transformation and a quadratic term of energy. We still find a positive and statistically significant relationship between ad-creative fixed effects and log(Energyi). The quadratic term, however, does not work because of the high correlation of .97 between the linear and the quadratic energy terms.

In terms of the magnitude of the estimate, we find that an increase of one standard deviation in an ad creative's energy level (.137) is associated with a $.137 \times 3.343 = .458$ percentage point increase in the ad-tuning rate. The number translates into about 5,079 more TVs tuned in for 75% or more of the ad's length per insertion (based on an average of 1,108,957 TVs tuned in per ad insertion in our data). Note that over 200 brands have within-brand energy level standard deviation greater than.137 (e.g., The North Face with.282, Apple Music with.246, and Under Armour with.246), which suggests that such changes in energy level are within the realm of possibility for advertisers. Overall, these results provide support for the positive association, namely, that more energetic ad content results in longer ad-tuning rates.

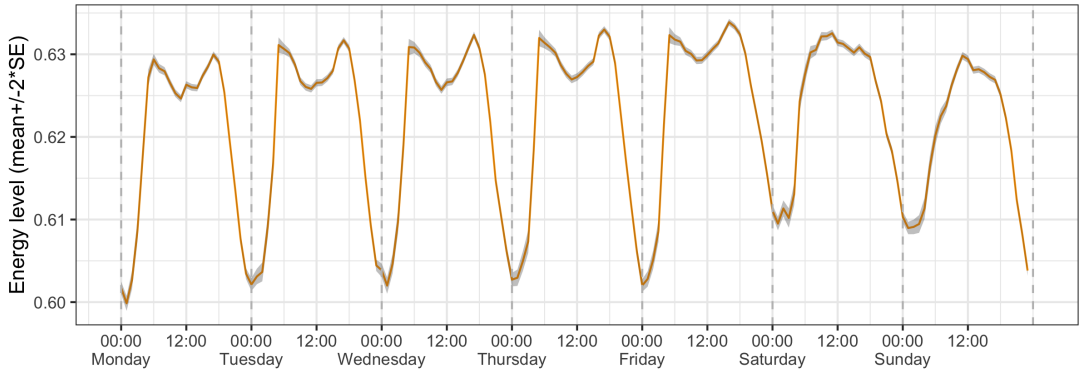
We also detect interesting patterns in the other coefficients. For instance, longer-duration ads have lower ad-creative fixed effects and hence would have lower ad-tuning rates. When compared with the baseline duration (i.e., 15-second ads), 90-second, 60-second, and 30-second ads, in that order, are less likely to be tuned in for more than 75% of their length, which makes intuitive sense. This pattern is also consistent with previous findings that longer-duration ads are more likely to be avoided (Elpers, Wedel, and Pieters (2003), p. 445). Our model-free raw data also show that the average ad-tuning rate systematically declines with increasing length of the ad. The results also indicate that (1) sales promotional ad creatives or ads with accompanying popular songs are less likely to be watched, (2) ad creatives that display an animal or animals are more likely to be watched, and (3) informational ad creatives are less likely to be watched than active ad creatives (active is the baseline for the mood variable). A detailed explanation of these findings is beyond the scope of this research.

3.5.2. A Within-Estimator Approach

The model results discussed previously relied on the between-ad variation in energy levels. That is, we compared ad A and ad B with different energy levels while controlling for the specificities of an insertion (via all the insertion fixed effects) and some other aspects of ad content. One concern regarding the between-estimator approach is potential omitted variable bias as one cannot test the sufficiency of a given set of control variables. Here, we propose an alternative estimator that relies on "artificial" within-ad variation in energy, which we generate by exploiting the temporal patterns in audience preference for high-energy audio.

The ideal variation should arise from randomizing the energy level of an ad creative while holding everything else fixed (i.e., insertion and other ad content). That is, we want to have *within-ad* variation in the energy level, where the source of variation is not correlated with other aspects of the ad. Using such variation, one can isolate the effect of energy levels in ads on outcome variables of interest from those held fixed. The challenge is that there is no such variation in the energy levels for a given ad creative due to the observational nature of our data. To partially address the problem, we rely on temporal variation in people's preferences for energetic audio. For instance, Park et al. (2019) report that people generally prefer more intense music during the day than at night. We also find the same pattern by examining the energy levels of 31 million listening sessions on Spotify (the MSSD), which represent 2.3 million tracks (or songs) played. A plot of the energy levels by time of day is shown in Figure 3.6.

Figure 3.6. Temporal patterns in the energy levels of MSSD dataset



Notes: The figure shows the temporal patterns in the auditory energy level of Spotify's Music Streaming Sessions Data (MSSD). This figure was generated from analyzing the energy levels of 31 million listening sessions of 2.3 million songs between July-September 2018. During the weekdays, energy level is relatively higher in daytime (6AM-6PM) than night; highest at 6AM and 6PM and lowest at midnight. During the weekends, the energy level is highest around noon.

This data set suggests a way of operationalizing the within-ad variation in energy levels. For a given ad creative on TV, the ad's energy level could be higher or lower than the preferred musical energy at the time when it is aired. By exploiting this natural occurrence, we can compare the ad-tuning rate of the same ad in the two cases: when the energy level of the ad is either higher or lower than the level generally preferred by people when the ad is aired. We consider the baseline energy level from the Spotify listening sessions in the MSSD (Figure 3.6) for each one-hour time slot over a 24-hour day of the seven days in the week as the preferred energy level. We compute the distance between the energy level of the particular ad creative shown in a specific one-hour time slot and the preferred baseline energy level for that time slot as computed from the MSSD data set. We relate the ad-tuning rate to the distance between an ad's energy level and the baseline level.

Specifically, we introduce a variable, $dist_{it}$, which measures the distance between ad creative i's energy level measured at the first 75% of an ad's length (Energy_i) and the baseline energy level $\theta_{\tau(t)}$ ($\tau(t)$ (the subscript $\tau(t)$ indicates the day-hour pair of t; e.g., Monday 1–2 P.M.). We cannot simply subtract the energy level of the TV ad creative from the energy level of the MSSD listening sessions because the level and variability of energy are different between the two. Therefore, we use deciles for both the energy level of ad creative i and the baseline energy level and compute the distance measure as $dist_{it} = \psi_i - \theta_{\tau(t)}$. Here, $\psi_i \in \{1, 2, ..., 10\}$ is the decile of ad creative i's energy level, and $\theta_{\tau(t)} \in \{1, 2, ..., 10\}$ is the decile group of the preferred energy level at $\tau(t)$, computed using the MSSD data set. Thus, $dist_{it}$ is one of the 19 integer values from -9 through 9; the more positive the value, the higher the relative energy level of ad creative i compared with the preferred baseline energy level.

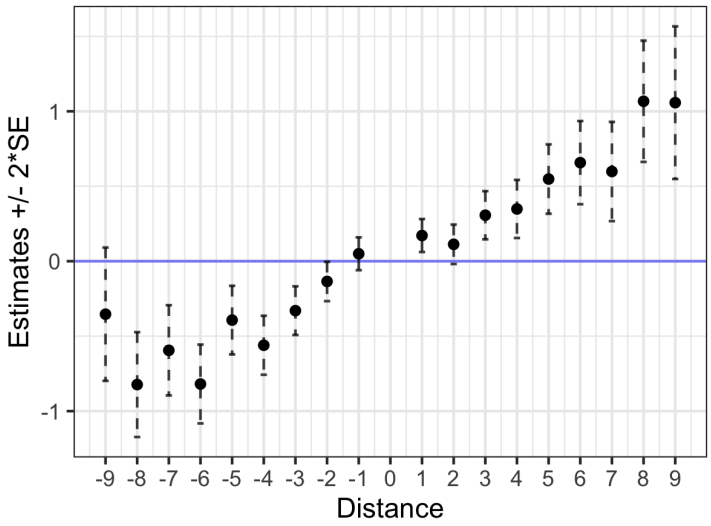
Using our distance measure, we estimate the following model:

$$\text{ATR}_{it} = \delta_i + X_t'\gamma + \sum_{d \in [-9,-8,...,8,9]} \beta^d \cdot I[dist_{it} = d] + \varepsilon_{it}, \tag{3.4}$$

where δ_i are the ad-creative fixed effects and X_t includes the same set of fixed effects as in Equation 3.2, namely network, program, year-week, day-of-week, day part, and position fixed effects. The terms β^d are the parameters of interest.

We graphically report the parameter estimates in Figure 3.7. We find that a relatively higher energy level than the baseline is associated with higher ad-tuning rates or lower ad avoidance. The association is almost monotonic across different values of distance. Reassuringly, these results provide support for the positive association, i.e., that higher energy in ad content is associated with longer ad-tuning rates.

Figure 3.7. Estimation results of the within estimator



Notes: The figure reports the parameter estimates of β^d in Equation 3.4.

An advantage of using the Spotify music energy baseline is that it represents individual preferences for audio energy over the day. There are, however, several caveats with this approach. First, the baseline is energy from Spotify music for a specific three-month period in 2018, whereas the ads in our data span a three-year period from 2015 to 2018. So, the temporal sample is not the same. Second, energy estimates for the MSSD listening tracks are directly provided by Spotify, whereas the energy estimates for TV ads were obtained with the machine learning procedure we described previously. Third, the baseline is music energy, but it is compared with ad energy, which is more than music. We cannot objectively confirm that this is a valid comparison, except to note that music is a major component of advertising. The fact remains, therefore, that Spotify listeners' preference for energetic music by itself may not be the same as the preferences of TV audiences and that the segments of TV audiences vary across different times.

To partially address these concerns, we use the aggregate energy level of ad creatives for a given time as the baseline.[15] Specifically, we compute the baseline as the average energy level of ad creatives on a particular network in a particular month, day of week, and hour of the day. The underlying assumption here is that TV audiences have a prior expectation of the energy level of ads during a given time, which we summarize by the average energy level of ad creatives. Also, advertisers select programs whose viewing demographics best match the product being advertised. Hence, one could expect average ad energy levels to reflect the desired energy levels. We find that an energy level higher than the baseline is positively associated with the ad-tuning rate, and the association is statistically significant, which supports our previous findings.[16] For more details, see Web Appendix E.1.

3.5.3. Interpretation

Overall, we find that the association between ad-tuning rate and energy levels in ads is positive and statistically significant using two empirical approaches. Nonetheless, we cannot claim that the association is causal, and we leave our findings as descriptive. This is because we cannot discount the possibility of an omitted variable bias as there could be missing variables that are related to the ad-creative effects (an example could be the message of ads). To the extent that these variables affect the ad-tuning rate and are also correlated with ad energy, our estimate is either under or overestimated. Further, given the observational nature of our data, the placement of ads with different energy levels

[15]We thank the review team for suggesting this idea.

[16]Relatedly, we also use the energy level of the previous ad within the same commercial break as a baseline. We find that the estimate from this approach is statistically insignificant.

is not randomized across programs and day parts, which together can create different contextual effects.

Relatedly, our reverse engineering of the Spotify algorithm using the FMA data set to measure energy in ad content could have generated a measurement error. The correlation between the predicted and the actual Spotify audio energy value available in the FMA data set is.89 in the holdout sample and.95 in the entire FMA data set. We find that the prediction is neither conservative nor extreme because most data points are aligned on the diagonal line without systematic bias (see Figure W1 in the Web Appendix). This measurement error could have biased the estimate toward zero, but we still find a positive and statistically significant association between energy in ad content and the average tuning rates. We expect that with a better prediction, for instance by having direct access to the Spotify algorithm, we would have been able to report a stronger association.

Lastly, because our energy measure is based on the auditory component of ads, it is worth noting that the association we report may be intertwined with energy levels in the visual content of ads. As discussed previously, however, the energy levels of visual and auditory components of ads are highly correlated at least from the standpoint of human perception. Since energetic audio and energetic video tend to go hand in hand, one could measure the energy level of ads using either audio or video or both. Thus, our estimates can be interpreted as the net effect of energy levels in both audio and video of ads on ad-tuning rates.

3.5.4. Heterogeneity

In the model results presented previously, we estimated the average association between energy and the quality of ad creatives in terms of ad-tuning rates across all product categories and insertions. Here, we analyze whether and how the association between energy in ad content and ad-tuning rates varies across product categories and airings. Ideally, we would like to estimate β in Equations 3.3 and 3.4 for any given ad creative i and insertion t. In practice, however, this approach is infeasible because each airing of an ad creative is unique and, hence, some degree of aggregation is necessary. Toward this end, we estimate the association for combinations of the three program genres and 15 product categories in the data. The first column of Table 3.5 reports the estimation results of Equation 3.3 in which we interact the energy variable with the product category dummies. The next three columns give results for each of the three program genres, thus allowing us to quantify the ad-creative fixed effects specifically for each program genre. We estimate the first stage in Equation 3.2 separately for each program genre to obtain genre-specific ad-creative fixed effects.

We find that the association of energy levels and ad-tuning rates is context dependent; that is, the magnitude and the statistical precision differ substantially across product categories and program genres. For instance, the association is statistically significant at the.05 level for only four product categories (business and legal, life and entertainment, travel, and vehicles) for all programs (Column 1). The number of categories with a significant relationship is even smaller when we consider a particular genre of programs: three for entertainment programs, one for news and information programs, and zero for sports programs. One concern is that lack of statistical significance simply arises from

the reduction in the sample size as we narrow our focus on a particular slice of the data. Consequently, we also check the results from a similar analysis using our within-estimator approach. In Table 3.6, we find that the association is statistically significant for more pairs of product categories and program genres, and some associations are even negative, although some estimates are different from Table 3.5, producing mixed results. Taken together, these results suggest that the boundary conditions for using high-energy features for TV commercials will include (1) what products are advertised and (2) the types of programs in which the ads are placed.

3.5.5. Additional Analyses

We briefly discuss additional robustness checks for our findings. More details on each analysis are reported in Web Appendixes E.2 and E.3.

Alternative definitions of the ad-tuning rate. Until now, we have used the 75% cutoff when defining the ad-tuning rate in the first stage of our between-estimator analysis. We have two additional cutoffs at 50% and 25% available in the data. We cannot vary the cutoff as the data are already aggregated. Using the 50% and 25% cutoffs, we find that the estimate of the energy effect is not significant. One explanation for the lack of significance of energy when the ad-tuning rate is measured for earlier cutoffs could be that arousal takes some time to build, so the shorter the tuning time, the lower the effect. We find that the results are consistent when we transform the dependent variable (either arcsine or logit) in the first stage to alleviate the concern regarding the normality assumption of the ad-tuning rate, which ranges between 0 and 100.

In addition, we vary the cutoff values depending on the length of the ads. For 15-second or shorter ads, we used a 75% cutoff. For ads with length of more than 15 seconds, up to 30 seconds, we used 50%. For ads longer than 30 seconds, we used 25%. This approach allows us to keep the duration of initial exposure to ads roughly comparable before viewers take any actions of avoidance. We make sure that the energy levels of the ads are also measured by taking into account the varying cutoff values (e.g., the first 75% to measure the energy level of a 15-second ad, or the first 50% to measure the energy level of a 30-second ad). We find that the results are qualitatively unchanged.

Accounting for ad-wearout effects. The intensity of previous exposure to the same ad is likely to be an important determinant of individuals' ad-avoidance behavior (often referred to as "ad-wearout effects"). As we do not have individual-level viewership data, which would have made accounting for such an effect relatively simple, we construct a variable that represents cumulative viewership prior to a specific airing of an ad and use it as an additional control in the first stage of our between-estimator analysis. Because we do not know when an ad was first aired, we split our three-year data into two periods (the first 12 months and the remaining 24 months), and use the first period to address the initial condition problem and the second period to reestimate the model. The estimate for the cumulative viewership variable is negative and statistically significant, which is consistent with the logic that ad repetition leads to lower ad-tuning rates. We also find that the ad-creative fixed effects obtained from the specification with or without controlling for ad-wearout effects are highly correlated (.998). The second-stage results are qualitatively unchanged even after controlling for ads' cumulative viewership in the first stage.

3.6. Conclusion

The research was motivated by the finding that the energy level in ads shown on TV has been increasing over time. Intrigued by this finding, we empirically investigate the relationship between the use of energetic ad content and the extent of ad-tuning time. A critical question is how energy is measured and what it represents. We estimate the energy level in ads by adapting a measure developed by Echo Nest, now owned by Spotify. To address what the measure represents, we validate the energy measure through an MTurk study. We find that the energy measure is strongly correlated with the construct of arousal. Arousal is directly linked to stimulation, which has been shown to affect ad tuning. The Spotify energy measure was estimated on music tracks, whereas our measure of energy in ads includes other background sounds, such as narration and product sounds, in addition to music. Further, visual aspects also contribute to the energy in an ad. We find through the MTurk study that the perceptions of audio energy and visual energy are strongly correlated, and that the Spotify energy measure is correlated to the perceived arousal of the audio part of the ad, the perceived arousal of the visual part of the ad, and the combined audio and visual energy of the ad. To summarize, the empirically estimated Spotify energy measure is linked to arousal and to both the audio and visual energy of the ads.

We evaluate the relationship between the energy levels in TV commercials and ad-tuning rate using a large-scale data set, while controlling for network effects, program effects, an extensive set of time effects, and position effects of the ad creative. We find that, on average, longer ad-tuning rates are associated with higher energy in ad content. At the same time, we find that the result is not uniform and depends on the program

genre and product categories. In the entertainment and news genres, we find a positive relationship between energy and ad-tuning rates in 11 of the 15 and 7 of the 15 product categories, respectively (Table 3.6; at the 10% significance level). A negative relationship is found in one product category in the news genre and three product categories in the sports genre.

Some findings in the literature, however, suggest that increases in energy may not always lead to increased viewing under certain conditions. Research by Puccinelli, Wilcox, and Grewal (2015), for instance, finds that highly energetic ads can lead to less watching when they are inserted in deactivating TV shows (e.g., sad movies). Craton and Lantos (2011) state that "music that is either overly or insufficiently arousing for a particular consumer in a specific context will be regarded unfavorably" (p. 405) and recommend that the level of stimulation provided by the music be matched to that of the program or ad content. In a follow-up study, Lantos and Craton (2012) point out that heterogeneity among viewers can lead to a positive or negative response to music in advertising, which can be extended to audio in general. These observations seem intuitive because judgments typically involve a reference to a standard, either internal or contextual. Thus, whether an ad is energetic might be evaluated in relation to the context. The implication is that the same level of arousal might be perceived as high in a low-arousal context but low in a high-arousal context. An important contextual factor in advertising therefore is the program in which the ad is run.

In real advertising settings, ads are seen in contexts with tremendous heterogeneity, which cannot be fully accounted for and controlled by advertisers. These observations may lead one to believe that the relationship between high energy in ads and the tendency

to view ads longer (or shorter) will always depend on context; hence, that generalizable patterns are difficult to discern. Although there is truth to this view, we ask in our research what the relationship looks like after controlling for network, program, time, and ad position. Advertisers typically cannot choose the ad position within a program, and in some cases the specific program in which the ad is shown as well depending on how ad time is purchased. However, advertisers select the network and the year-week, day of the week, and hour of the day when their ads are shown. Our results suggest that, overall, there is a positive association between energy levels in ad content and one aspect of viewers' behavior, namely, ad tuning. We find that the association, on average, is positive, but it varies both in magnitude and in direction across pairs of a program genre and a product category.

We also note that ad tuning may be affected by the presence and timing of when different objects appear in the ad. For instance, Teixeira, Wedel, and Pieters (2010) find that the timing of the appearance of the advertised brand's logo in an ad affects ad-tuning rates. Similarly, consumer reaction to ads can be affected by whether ads include animals (Trivedi and Teichert, 2020). Our focus in this research, however, is on the role of an ad's energy in tuning behavior. Next, we discuss managerial implications of our findings.

3.6.1. Managerial Implications

Advertisers and television networks have routinely included audio in ads that is much louder than the audio in the programs in which they are aired. The assumption has been that an increase in the loudness of the audio of ads relative to that in the program attracts attention to the ads. Increasing attention to ads has been linked to lower ad

avoidance (Teixeira, Wedel, and Pieters (2012); Tse and Lee (2001)). The practice of making ads louder than the programs they are aired in became so prevalent that it raised concerns about the health effects of loudness on viewing audiences, leading to regulatory limitations on how much louder ads can be than the programs in which they are placed. The resulting CALM (Commercial Advertisement Loudness Mitigation) Act, which was passed in 2010, and began to be enforced by the Federal Communications Commission in 2012, limits the average loudness of an ad to no more than the average loudness of the program in which it is aired. Advertisers and networks, therefore, cannot continue to rely on loudness as a means of attracting attention to reduce ad avoidance. In addition, television manufacturers and streaming service providers have begun to offer features that give television audiences more control over the loudness of the programming. In principle, audiences can either mute ads or force them to be at the same level of sound as the programs they are aired in. Advertisers, therefore, need to be creative in how to use audio in their ads to attract and retain audience attention. This is a particularly vexing challenge because of the increasing problem of ads being avoided and skipped by viewers. Our results suggest that, on average, increasing the energy in an ad can increase ad tuning or reduce ad avoidance. This finding is consistent with that of Belanche, Flavián, and Pérez-Rueda (2017). We also note that energy is not just loudness, and ad energy comprises both audio energy and visual energy. In our MTurk study, participants associated energetic ads with descriptors like "fast," "music," "movement," "upbeat," and "exciting." When testing different creatives, therefore, advertisers should also focus on evaluating the effects of different aspects of energy in the ability of ad creatives to gain higher viewing rates. Given our findings of heterogeneity, advertisers will also need

to experiment and identify the program genres in which ad creatives with specific levels of energy are likely to be successful in attracting longer viewing rates.

Our empirical measure of energy includes all of the audio in the ad, which includes sounds associated with the visual aspects of the ad. A useful next step therefore would be to map levels of energy in ad content into a granular feature set, which includes both auditory components (e.g., dynamic range, entropy, loudness, onset rate, and timbre) and visual components (e.g., colorfulness, saturation, human facial expressions, motion, scene similarity). If a creative is being designed with background music, those composing the music should use all of these attributes to reach the target level of energy for the creative. For instance, they could increase the dynamic range of the music. Alternatively, they could increase the music's entropy, onset rate, or timbre to reach the target level. The same could be done with any speech or other visual elements in the creative. Then, the next step would be understanding how each component contributes to a certain aspect of consumer behavior, as well as how multiple components interact. This step would require data with (ideally exogenous) variation in each component. For instance, one could design an experiment in which the energy levels in audio and video of ads are randomized to tease out the relative role of the two variations, which we lack in our data.

Such strategic determination of ad content can benefit ad creators, as well as advertising publishers. Ad creators working with objective algorithms (e.g., MIR software) should be able to measure the values of these attributes for any auditory composition. They can then adjust the attributes to reach the targeted audio energy levels. Digital ad publishers, such as Spotify and Pandora, can also leverage the auditory attributes of

tracks in users' listening sessions to better target ads. They could also inform advertisers to better design ad creatives based on the advertisers' target listeners.

3.6.2. Limitations and Directions for Future Research

The data used in this research are based on actual observations of the second-to-second tuning behaviors across millions of televisions rather than the smaller samples and specialized settings used in other studies of ad viewing behaviors. Several limitations, however, are in order. First, because of the observational nature of our data, our findings cannot be construed as capturing causal relationships. They are only correlational. Also, this study does not attempt to reveal the mechanism under which ad content affects consumer behavior. Second, we do not know whether the individual or individuals tuned in to the monitored televisions were indeed viewing or paying attention to the aired ads when the TV sets were tracked by our data provider. Relatedly, we have no data on other metrics that advertisers are interested in knowing, such as likability, memorability, recall, and conversion to purchase. Third, our data are missing consumer behaviors from time-shifted views, which are reported to be growing. Audience behaviors regarding ad avoidance may be different between live versus time-shifted views. Fourth, our data lack independent variation in the energy levels in ad audio and video, which prevents us from investigating any meaningful interaction between the two. Also, we focus on the energy extracted from overall sound in ads rather than distinguishing between human voice and background sound. Lastly, our focus is on the overall association between energy in ad content and ad-tuning rate. As mentioned, however, in certain cases higher energy can lead to lower ad-tuning rates. Our results also show that the association varies across

program genres and product categories, but we are unable to reveal the mechanism or mechanisms that explain when and how high energy in ad content is effective.

Although these limitations each point to a direction for future research on how the content of ads can influence their effectiveness, we believe an important next step is to further investigate our findings through controlled large-scale testing in a real-world context. We hope that this study motivates the initiation of such testing and provides initial guidelines for the design of such studies.

Table 3.4. Between-Estimator Approach: Second-Stage Estimation Results

	DV: Estimated Ad-Creative Fixed Effects ($\hat{\delta}_i$)			
	(1)	(2)	(3)	(4)
Energy	.928	1.094	1.180	3.343***
	(.811)	(.802)	(.811)	(.922)
Duration: 30 seconds		-2.790***	-2.877***	-1.806***
		(.228)	(.229)	(.278)
Duration: 60 seconds		-6.308***	-5.939***	-3.796***
		(.546)	(.560)	(.789)
Duration: 90 seconds		-3.872**	-2.378	-10.680***
		(1.657)	(1.674)	(2.710)
Duration: >90 seconds		-21.842***	-20.998***	-21.305***
		(1.774)	(1.789)	(2.317)
Promotion		-5.033***	-4.954***	-2.401***
		(.258)	(.259)	(.369)
Animal		1.841***	1.678***	.940**
		(.389)	(.390)	(.479)
Song		-1.018***	-1.151***	-.647*
		(.298)	(.300)	(.357)
Mood: Emotional			1.427**	.465
			(.673)	(.791)
Mood: Informational			-2.105***	-1.056*
			(0.518)	(0.615)
Mood: Funny			.589**	.520
			(.278)	(.342)
Mood: Sexy			-.271	-1.848
			(1.279)	(1.850)
Constant	81.103***	83.928***	83.859***	
	(0.284)	(0.311)	(0.322)	
Brand fixed effects	No	No	No	Yes
N	18,933	18,923	18,861	18,861
R^2	<.001	.042	.044	.257
$Adj.R^2$	<.001	.041	.043	.152

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Note: The table reports the estimation results of Equation 3.3; standard errors are reported in parentheses.

Table 3.5. Heterogeneous Associations from the Between-Estimator Approach

	DV: Estimated Ad-Creative Fixed Effects ($\hat{\delta}_{i\|g}$)			
	(1) All Programs	(2) Entertainment	(3) News & Information	(4) Sports
Energy				
× Apparel, footwear, accessories	-5.821	-8.921	-12.384	-6.754
	(5.275)	(7.257)	(10.051)	(8.591)
× Business and legal	8.719**	1.070	7.114	9.659
	(4.153)	(5.350)	(7.167)	(6.905)
× Education	3.073	-14.060	0.229	8.693
	(12.366)	(17.152)	(17.303)	(37.682)
× Electronics and communication	-0.351	-1.004	7.775	-4.210
	(2.280)	(2.593)	(5.065)	(3.800)
× Food and beverage	2.908	5.549**	2.995	-4.740
	(2.200)	(2.476)	(3.353)	(4.928)
× Health and beauty	0.419	1.106	-1.842	5.824
	(2.741)	(2.864)	(3.611)	(10.417)
× Home and real estate	1.242	1.480	-4.657	-3.478
	(3.677)	(4.007)	(4.936)	(11.534)
× Insurance	8.608*	8.050	9.791	1.587
	(4.467)	(5.062)	(6.003)	(6.626)
× Life and entertainment	9.668**	9.580*	10.966	-4.065
	(3.963)	(5.160)	(11.181)	(7.953)
× Pharmaceutical and medical	5.750	9.558*	10.976**	-16.058
	(5.227)	(5.393)	(5.616)	(23.303)
× Politics, government, organizations	11.072	10.843	11.483	15.715
	(9.275)	(12.464)	(11.993)	(22.582)
× Restaurants	-0.274	-1.489	2.516	-0.014
	(3.625)	(4.031)	(7.321)	(6.241)
× Retail stores	4.206	2.508	2.976	-18.993
	(2.718)	(2.954)	(3.890)	(10.702)
× Travel	25.371***	34.235***	21.059	-18.145
	(8.165)	(9.879)	(14.016)	(17.568)
× Vehicles	7.050**	11.210***	-10.152	0.346
	(3.219)	(4.006)	(7.266)	(5.212)
Other controls	Yes	Yes	Yes	Yes
Brand fixed effects	Yes	Yes	Yes	Yes
N	18,861	14,637	7,537	4,934
R^2	0.259	0.268	0.295	0.281
$Adj.R^2$	0.152	0.163	0.148	0.104

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Note: The table reports the estimation results of Equation 3.3 using FGLS by product categories and program genres. The sum of the sample sizes in each column exceeds the number of ad creatives because the same ad creative can be shown in entertainment, news and information, and sports genres.

Table 3.6. Heterogeneous Associations from the Within-Estimator Approach

	DV: Ad Tuning-Rate (ATR_{it})			
	1) All Programs	(2) Entertainment	(3) News & Information	(4) Sports
Energy				
× Apparel, footwear, accessories	-.060	-.010	.230**	-.253*
	(.042)	(.101)	(.110)	(.131)
× Business and legal	-.012	.133**	.092	.097
	(.030)	(.065)	(.057)	(.103)
× Education	-.003	-.072	-.086	.129
	(.078)	(.117)	(.192)	(.373)
× Electronics and communication	.112***	.125***	.092	.066
	(.023)	(.040)	(.066)	(.075)
× Food and beverage	.061***	.242***	.092***	-.129
	(.020)	(.029)	(.035)	(.083)
× Health and beauty	.101***	.281***	.105***	-.296**
	(.021)	(.029)	(.035)	(.138)
× Home and real estate	-.001	.170***	-.112**	.306*
	(.025)	(.032)	(.048)	(.165)
× Insurance	.119***	.072*	.064	-.157*
	(.025)	(.040)	(.058)	(.086)
× Life and entertainment	.060	.118	.167	.044
	(.041)	(.075)	(.107)	(.118)
× Pharmaceutical and medical	.045**	.234***	-.014	-.137
	(.022)	(.031)	(.035)	(.174)
× Politics, government, organizations	.158**	.234*	.244**	.018
	(.062)	(.135)	(.123)	(.205)
× Restaurants	.060*	.025	-.128	.177*
	(.032)	(.053)	(.098)	(.102)
× Retail stores	.113***	.139***	.145***	-.006
	(.026)	(.038)	(.054)	(.138)
× Travel	.166***	.142*	.350***	.521***
	(.045)	(.082)	(.092)	(.174)
× Vehicles	.085***	.145**	.255***	.248***
	(.028)	(.057)	(.082)	(.087)
Ad-creative fixed effects	Yes	Yes	Yes	Yes
Insertion fixed effects[a]	Yes	Yes	Yes	Yes
N	1,057,798	728,246	237,267	83,484
R^2	.638	.653	.641	.711
$Adj.R^2$	.631	.645	.621	.675

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Note: The table reports the estimation results of Equation 3.4 by product categories and program genres. For the ease of interpretation, the variable $dist_{it}$ is entered in the equation linearly.

Bibliography

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.

Askin, Noah and Michael Mauskapf (2017). "What makes popular culture popular? Product features and optimal differentiation in music". *American Sociological Review* 82.5, 910–944.

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.

Balducci, Bitty and Detelina Marinova (2018). "Unstructured data in marketing". *Journal of the Academy of Marketing Science* 46.4, 557–590.

Belanche, Daniel, Carlos Flavián, and Alfredo Pérez-Rueda (2017). "Understanding interactive online advertising: Congruence and product involvement in highly and lowly arousing, skippable video ads". *Journal of Interactive Marketing* 37, 75–88.

Bellotti, Victoria et al. (2017). "Why Users Disintermediate Peer-to-Peer Marketplaces". In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 4370–4382. ISBN: 9781450346559.

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

Boughanmi, Khaled and Asim Ansari (2021). "Dynamics of musical success: a machine learning approach for multimedia data Fusion". *Journal of Marketing Research* 58.6, 1034–1057.

Brost, Brian, Rishabh Mehrotra, and Tristan Jehan (2019). "The music streaming sessions dataset". In *The World Wide Web Conference*, 2594–2600.

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.

Craton, Lincoln G and Geoffrey P Lantos (2011). "Attitude toward the advertising music: An overlooked potential pitfall in commercials". *Journal of Consumer Marketing* 28.6, 396–411.

Cusumano, Michael A., Annabelle Gawer, and David B. Yoffie (2020). "The Future of Platforms". *MIT Sloan Management Review* 61.3, 26–34.

Danaher, Peter J (1995). "What happens to television ratings during commercial breaks?" *Journal of advertising Research* 35.1, 37–37.

Defferrard, Michaël et al. (2017). "FMA: A Dataset for Music Analysis". In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*. Ed. by Sally Jo Cunningham et al., 316–323.

Deng, Yiting and Carl F Mela (2018). "TV viewing and advertising targeting". *Journal of Marketing Research* 55.1, 99–118.

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *CoRR* abs/1810.04805.

Di Muro, Fabrizio and Kyle B Murray (2012). "An arousal regulation explanation of mood effects on consumer choice". *Journal of Consumer Research* 39.3, 574–584.

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

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

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

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.

Elpers, Josephine LCM Woltman, Michel Wedel, and Rik GM Pieters (2003). "Why do consumers stop viewing television commercials? Two experiments on the influence of moment-to-moment entertainment and information value". *Journal of Marketing Research* 40.4, 437–453.

Ferraro, Andres, Dmitry Bogdanov, and Xavier Serra (2019). "Skip prediction using boosting trees based on acoustic features of tracks in sessions". *arXiv preprint arXiv:1903.11833*.

Fong, Hortense, Vineet Kumar, and K Sudhir (2021). "A theory-based interpretable deep learning architecture for music emotion". *Available at SSRN 4025386*.

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*.

Gorn, Gerald, Michel Tuan Pham, and Leo Yatming Sin (2001). "When arousal influences ad evaluation and valence does not (and vice versa)". *Journal of consumer Psychology* 11.1, 43–55.

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`.

— (2023). "Marketplace leakage". *Management Science*.

He, Eryn Juan et al. (2020). "Off-platform Threats in On-Demand Services". *SSRN 3550646*.

Jalali, Nima Y and Purushottam Papatla (2016). "The palette that stands out: Color compositions of online curated visual UGC that attracts higher consumer interaction". *Quantitative Marketing and Economics* 14.4, 353–384.

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 (2021). "The Achilles heel of the platform business model: Disintermediation". *Business Horizons*.

Lantos, Geoffrey P and Lincoln G Craton (2012). "A model of consumer response to advertising music". *Journal of Consumer Marketing* 29.1, 22–42.

Lewis, Jeffrey B and Drew A Linzer (2005). "Estimating regression models in which the dependent variable is based on estimates". *Political analysis* 13.4, 345–364.

Li, Xi, Mengze Shi, and Xin Shane Wang (2019). "Video mining: Measuring visual information using automatic methods". *International Journal of Research in Marketing* 36.2, 216–231.

Liu, Liu, Daria Dzyabura, and Natalie Mizik (2020). "Visual listening in: Extracting brand image portrayed on social media". *Marketing Science* 39.4, 669–686.

Liu, Xiao, Param Vir Singh, and Kannan Srinivasan (2016). "A structured analysis of unstructured big data by leveraging cloud computing". *Marketing Science* 35.3, 363–388.

Lu, Shasha, Li Xiao, and Min Ding (2016). "A video-based automated recommender (VAR) system for garments". *Marketing Science* 35.3, 484–510.

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`.

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.

Marinova, Detelina, Sunil K Singh, and Jagdip Singh (2018). "Frontline problem-solving effectiveness: A dynamic analysis of verbal and nonverbal cues". *Journal of Marketing Research* 55.2, 178–192.

McFadden, Daniel and Kenneth Train (2000). "Mixed MNL models for discrete response". *Journal of applied Econometrics* 15.5, 447–470.

McFee, Brian et al. (2015). "librosa: Audio and music signal analysis in python". In *Proceedings of the 14th python in science conference*. Vol. 8, 18–25.

McGranaghan, Matthew, Jura Liaukonyte, and Kenneth C Wilbur (2022). "How viewer tuning, presence, and attention respond to ad content and predict brand search lift". *Marketing Science* 41.5, 873–895.

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.

Mehrabian, Albert and James A Russell (1974). *An approach to environmental psychology*. the MIT Press.

Nash, John F (1950). "The bargaining problem". *Econometrica: Journal of the econometric society*, 155–162.

Netzer, Oded et al. (2012). "Mine your own business: Market-structure surveillance through text mining". *Marketing Science* 31.3, 521–543.

Ng, Aaron and Rishabh Mehrotra (2020). "Investigating the Impact of Audio States & Transitions for Track Sequencing in Music Streaming Sessions". In *Proceedings of the 14th ACM Conference on Recommender Systems*, 697–702.

Olney, Thomas J, Morris B Holbrook, and Rajeev Batra (1991). "Consumer responses to advertising: The effects of ad content, emotions, and attitude toward the ad on viewing time". *Journal of consumer research* 17.4, 440–453.

Park, Minsu et al. (2019). "Global music streaming data reveal diurnal and seasonal patterns of affective preference". *Nature Human Behaviour* 3.3, 230–236.

Peitz, Martin and Anton Sobolev (2022). "Inflated recommendations". CEPR Discussion Paper No. DP17260.

Preston, Ivan L (1982). "The association model of the advertising communication process". *Journal of Advertising* 11.2, 3–15.

Puccinelli, Nancy M, Keith Wilcox, and Dhruv Grewal (2015). "Consumers? response to commercials: when the energy level in the commercial conflicts with the media context". *Journal of Marketing* 79.2, 1–18.

Rochet, Jean-Charles and Jean Tirole (2006). "Two-Sided Markets: A Progress Report". *The RAND Journal of Economics* 37.3, 645–667.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2020). "A Primer in BERTology: What we know about how BERT works". *CoRR* abs/2002.12327.

Rubinstein, Ariel (1982). "Perfect equilibrium in a bargaining model". *Econometrica: Journal of the Econometric Society*, 97–109.

Russell, James A (1980). "A circumplex model of affect." *Journal of personality and social psychology* 39.6, 1161.

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

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

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`.

Siddarth, Sivaramakrishnan and Amitava Chattopadhyay (1998). "To zap or not to zap: A study of the determinants of channel switching during commercials". *Marketing Science* 17.2, 124–138.

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.

Teixeira, Thales, Rosalind Picard, and Rana El Kaliouby (2014). "Why, when, and how much to entertain consumers in advertisements? A web-based facial tracking field study". *Marketing Science* 33.6, 809–827.

Teixeira, Thales, Michel Wedel, and Rik Pieters (2012). "Emotion-induced engagement in internet video advertisements". *Journal of marketing research* 49.2, 144–159.

Teixeira, Thales S, Michel Wedel, and Rik Pieters (2010). "Moment-to-moment optimal branding in TV commercials: Preventing avoidance by pulsing". *Marketing Science* 29.5, 783–804.

Trivedi, Rohit H and Thorsten Teichert (2020). "Consumer Reactions to Animal And Human Models in Print Ads: How Animals and People in Ads Influence the Purchase-Decision Journey". *Journal of Advertising Research* 60.4, 426–438.

Tse, Alan Ching Biu and Ruby PW Lee (2001). "Zapping behavior during commercial breaks". *Journal of Advertising Research* 41.3, 25–29.

Tuchman, Anna E, Harikesh S Nair, and Pedro M Gardete (2018). "Television ad-skipping, consumption complementarities and the consumer demand for advertising". *Quantitative Marketing and Economics* 16.2, 111–174.

Uber (2022). *Uber prohibits Street Hails and Off-platform Pickups*. URL: `https://www.uber.com/legal/en/document/?name=general-community-guidelines&country=united-states&lang=en`.

Van Meurs, Lex (1998). "Zapp! A study on switching behavior during commercial breaks". *Journal of Advertising Research* 38.1, 43–44.

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.

Wilbur, Kenneth C (2008). "How the digital video recorder (DVR) changes traditional television advertising". *Journal of Advertising* 37.1, 143–149.

— (2016). "Advertising content and television advertising avoidance". *Journal of Media Economics* 29.2, 51–72.

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". In *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.

Xiao, Li and Min Ding (2014). "Just the faces: Exploring the effects of facial features in print advertising". *Marketing Science* 33.3, 338–352.

Xiao, Li, Hye-Jin Kim, and Min Ding (2013). "An introduction to audio and visual research and applications in marketing". *Review of Marketing Research* 10, 213–253.

Xie, Yingkang, Jingyi Wang, and Huaiyu Zhu (2022). "Monitoring Disintermediation: Actions Matter More Than Words". KDD'22 Workshop on Decision Intelligence and Analytics for Online Marketplaces: Jobs, Ridesharing, Retail, and Beyond.

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

Appendix for Chapter 2 on Platform Leakage

Appendix

Platform Leakage:
Incentive Conflicts in Two-Sided Markets

Yingkang Xie* Huaiyu Zhu

Table of Contents

"These materials have been supplied by the authors to aid in the understanding of their paper."

* Xie: doctoral student at the Kellogg School of Management, Northwestern University, Evanston, IL 60201 (yingkang.xie@kellogg.northwestern.edu); Zhu: senior director at Lalamove (zhuhuaiyu@gmail.com).

A.1. 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[1] [2] [3]. 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 A.1).

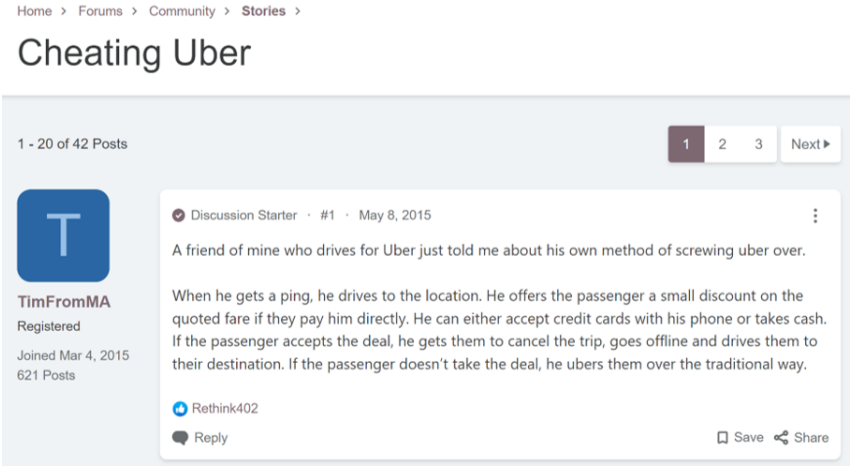
However, the extent of leakage varies widely across marketplaces. A higher price increases the absolute size of the commission (even if it is a low percentage), which raises the

[1]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

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

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

Figure A.1. 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/`)

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 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.

A.2. 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.

A.2.1. 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[4] 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, 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.

A.2.2. Keyword Detection (Indirect Measure)

Monitoring conversations is a common and established practice (Airbnb, 2022b; eBay, 2022a; 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

[4]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.

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.

A.2.3. 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 A.1. 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 A.1. 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).

A.3. Synthetic Control Methods (SCM)

A.3.1. The SCM Estimates of Beijing

We estimate the city-specific treatment effects on Non-VIPs using synthetic controls for 33 cities in Section A.3.2. 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 A.4.1, 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:

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

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 A.2 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:

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

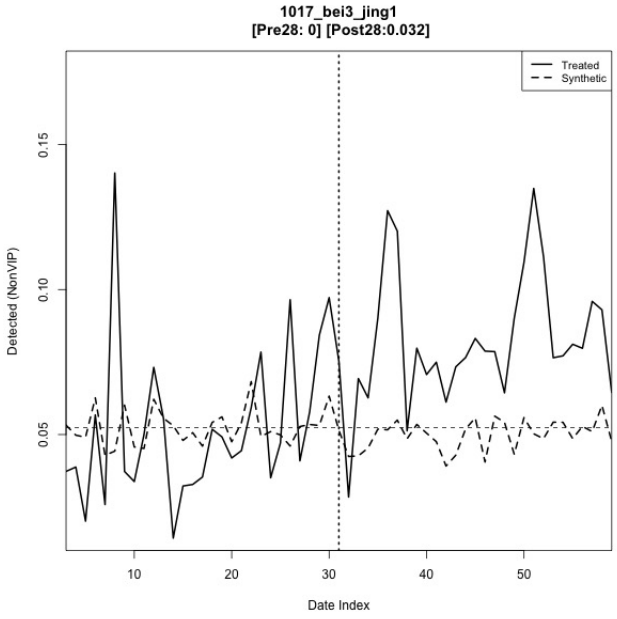


Figure A.2. Detected Rate

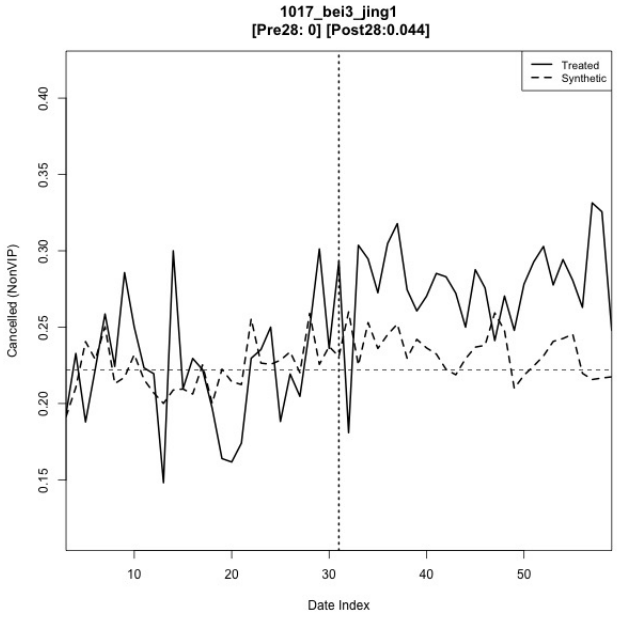


Figure A.3. Cancellation Rate

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

A.2 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 A.2) 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 A.3)after charging the commission fee.

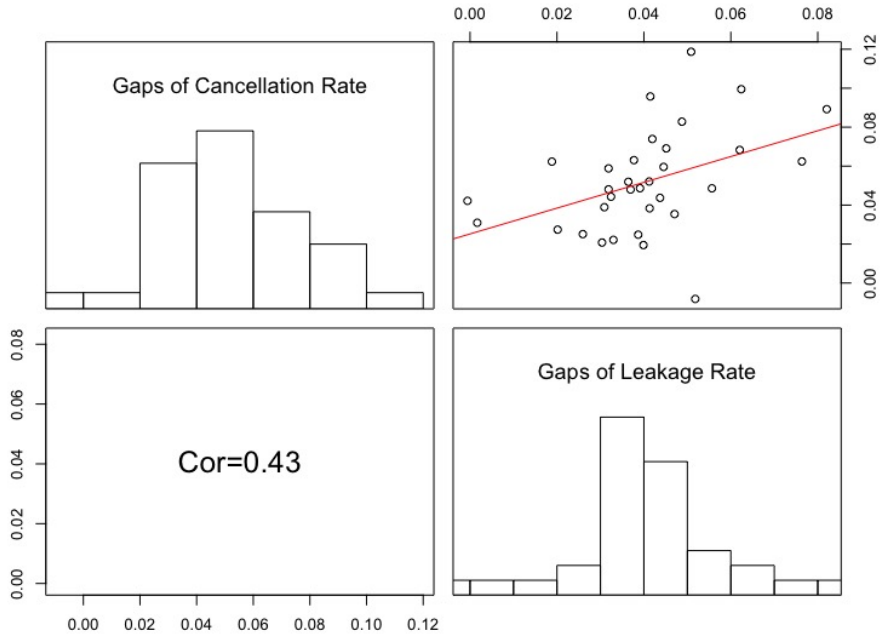
A.3.2. The Distribution of SCM Estimates of All Treated Cities

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 2.4. 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 (A.1) for all 33 cities that are treated in the 137 days. Figure A.4 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 (A.2) in the post-period. The top right panel shows that treatment effects on the two metrics are positively correlated and mostly non-negative.

We can conduct similar exercise for Super-VIPs. Table A.2 documents the summary statistics of city-specific treatment effects for both the Non-VIPs and Super-VIPs in the

Figure A.4. The City-Specific Treatment Effects for Non-VIPs from Synthetic Control Method



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 A.4) 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 A.6 than the ones in Figure A.5.

Table A.2. 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 A.5 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 A.6 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.

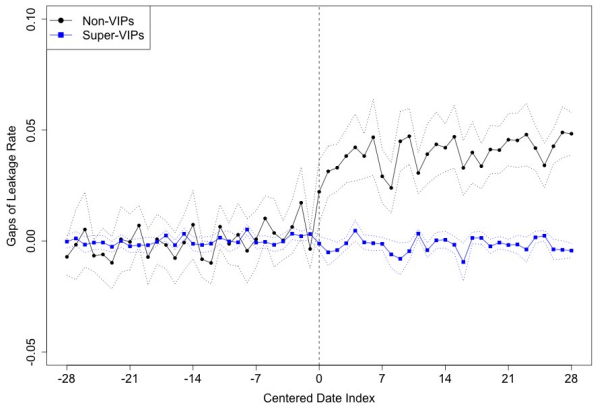


Figure A.5. Effects on Detected Rate

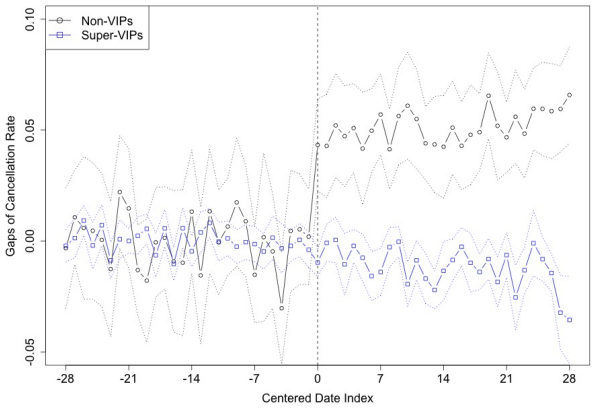


Figure A.6. Effects on Cancellation Rate

The histograms of estimates in Figure A.4 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 2.4.

A.4. Difference-in-Difference (DiD)

In Section A.3.1, 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.

A.4.1. 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[5] introduction of commission is exogenous because different VIP expiration dates shift the incentive shock across drivers in the same city.

[5]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.

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:

$$\text{(A.3)}\quad Y_{ict} = \alpha_i + \beta \cdot X_{ict} + \beta_v \cdot NonVIP_{ict} + \beta_p \cdot PolicyV2_{ict} + \delta \cdot PolicyV2_{ict} \times NonVIP_{ict} + \epsilon_{ict}$$

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 A.3 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 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.

Table A.3. 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

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.

A.4.2. Discussion: Suggestive vs. Causal Evidence

Both approaches in Section A.3.1 (synthetic control unit at the city level) and Section A.4.1 (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.

A.5. 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

$$L_{i(t)} = 1 \quad \text{iff} \quad \gamma_i p_t - \lambda_{i(t)} p_t > h_{i(t)} \tag{A.4}$$

Equation (A.4) 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 (2.3) in Section 2.4.2.1. 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 intuition of identification.

Data Points and Identification

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

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

$$L_{i(t)} = 0 \quad \text{iff} \quad 0 - \lambda_{i(t)} \leq h_{i(t)} \tag{A.5}$$

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

$$L_{i(t')} = 0 \quad \text{iff} \quad 0.15 p_{t'} - \lambda_{i(t')} > h_{i(t')} \tag{A.6}$$

(3) Leakage on transaction t'' with $\gamma_{it''} = 0.15$

$$L_{i(t'')} = 1 \quad \text{iff} \quad 0.15p_{t''} - \lambda_{i(t'')} > h_{i(t'')} \tag{A.7}$$

First, let's focus the hassle for driver i in transaction t. 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(A.6) and Eq(A.7): $h_i \in [0.15p_{t'}, 0.15p_{t''}]$
- Within-unit shock variation: same driver seeing same price before and after
  - Eq(A.5) and Eq(A.6): $\lambda_{i)} \in [\frac{0.15p_t - 2h_i}{2p_t}, \infty]$
  - Eq(A.5) and Eq(A.7): $\lambda_{i)} \in [-\frac{h_i}{p_t}, 0.15 - \frac{h_i}{p_t}]$
- Within-unit price and shock variation: different prices before and after the shock
  - Eq(A.5) and Eq(A.6): $\lambda_{i)} \in [\frac{0.15p_{t'} - 2h_i}{p_t + p_{t'}}, \infty]$
  - Eq(A.5) and Eq(A.7): $\lambda_{i)} \in [0, \min\{0.15 - \frac{h_i}{p_{t''}}, \frac{0.15p_{t''} - 2h_i}{p_{t''} - p_t}\}]$
  - Eq(A.6) and Eq(A.7): $\lambda_{i)} \in [0.15 - \frac{h_i}{p'_t}, 0.15 - \frac{h_i}{p''_t}]$

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.

A.6. Additional Tables of Structural MLE Estimates

Table A.4. 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)
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 A.5. 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

A.7. Microfounded Bargaining

A.7.1. Nash Bargaining

Eq (2.8) 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:

$$\begin{aligned} \Delta\pi_{ij(t)} &= \Pi^1_{i(t)} - \Pi^0_{i(t)} \\ &= \beta_i \cdot (\gamma_{i(t)} - \lambda_{ij(t)})p_t - h_{i(t)} \end{aligned} \tag{A.8}$$

and

$$\begin{aligned} \Delta u_{ji(t)} &= U^1_{j(t)} - U^0_{j(t)} \\ &= \beta_j \cdot (\lambda_{ij(t)}p_t - s_{j(t)}) - h_{j(t)} \end{aligned} \tag{A.9}$$

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

$$\begin{aligned} &\max_{\lambda_{ij(t)}} (\Delta\pi_{ij(t)})^{\eta} \cdot (\Delta u_{ji(t)})^{1-\eta} \\ &\text{s.t.} \quad \Delta\pi_{ij(t)} \geq 0 \\ &\qquad \Delta u_{ji(t)} \geq 0 \\ &\qquad 0 \leq \eta \leq 1 \end{aligned} \tag{A.10}$$

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:

$$\frac{\eta}{1-\eta} = \frac{\Delta\pi_{ij(t)}}{\Delta u_{ji(t)}} \tag{A.11}$$

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

$$\lambda^*_{ij(t)} = \frac{(1-\eta)\beta_i\gamma_{i(t)} + \eta\beta_j\frac{s_{j(t)}}{p_t} + \eta\frac{h_{j(t)}}{p_t} - (1-\eta)\frac{h_{i(t)}}{p_t}}{(1-\eta)\beta_i + \eta\beta_j} \tag{A.12}$$

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.

A.7.2. Estimation and Simulations

We can plug the analytical solution of Equation (A.12) back into Equation (2.8) to simulate the choice of leakage. We use the model specification in Section 2.5.4 for estimation.

(A.13)

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

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.

(A.14)

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

Equation (A.14) 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 (A.13) 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 A.4 show that experienced users who have repeated transactions on the platform can obtain an offline discount ranged from 16% to 21.6% (see Section 2.5.3.2). We will discuss what determines bargaining power in Appendix A.7.3.

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 2.10) using Equation (A.13). 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 ($\eta = 0$).

A.7.3. 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.

APPENDIX B

Appendix for Chapter 3 on Ad Avoidance

Web Appendix

High-Energy Ad Content:
A Large-scale Investigation of TV Commercials

Joonhyuk Yang† Yingkang Xie*
Lakshman Krishnamurthi Purushottam Papatla

Table of Contents

"These materials have been supplied by the authors to aid in the understanding of their paper."

† The first two authors contributed equally to this work in *Journal of Marketing Research* 59.4 (2022): 840-859.
* Yang: Assistant Professor of Marketing at Mendoza College of Business, University of Notre Dame (joonhyuk.yang@nd.edu); Xie: doctoral student at Kellogg School of Management, Northwestern University (yingkang.xie@kellogg.northwestern.edu); Krishnamurthi: A. Montgomery Ward Professor of Marketing at Kellogg School of Management (laksh@kellogg.northwestern.edu); Papatla: Northwestern Mutual Data Science Institute Professor at Lubar School of Business, University of Wisconsin-Milwaukee (papatla@uwm.edu).

WA-A. AUDIO ENERGY EXTRACTION AND PREDICTION

We provide a general framework to extract auditory characteristics from ad content.

A.1.1 Echo Nest Attributes. The Echo Nest audio attributes mentioned in the paper are considered "the current gold standard in music information retrieval (MIR)" (Askin and Mauskapf 2017). Using web crawling and audio encoding technologies, the Echo Nest Lab created objective qualities of audio recordings that intuitively describe music. In March 2014, The Echo Nest Lab was acquired by Spotify to power its analytics and recommendation engines. Before the acquisition, the company created Application Programming Interfaces (APIs) that were used by over 7,000 developers and powered more than 400 apps and sites. After the acquisition, The Echo Nest API was shut down on May 31, 2016. Developers were encouraged to use the Spotify API instead to obtain the Echo Nest attributes for over 50 million tracks available in the music streaming service.[1] The Echo Nest attributes have been widely adopted in the music industry. Leading music services (e.g., Clear Channel's iHeartradio, MOG, Rdio, SiriusXM, Spotify, etc.), editorial, video and social media networks (e.g., BBC.com, Foursquare, MTV, Twitter, VEVO, Yahoo!, etc.), connected device manufacturers such as doubleTwist, Nokia, etc., and big brands (e.g., Coca Cola, Intel, Microsoft, Reebok, etc.) use the platform and solutions to better understand music content. For example, Spotify uses the algorithms to power product features such as Sort Your Music[2] to engage their consumers.

The Echo Nest attributes have also been used in research for music preferences and recommendations. Park et al. (2019) used the Echo Nest attributes to demonstrate the affective preferences for music across user groups and countries. Millecamp et al. (2018) investigated the design of visual controls by users, such as the use of radar charts or sliders, to change music attributes that allow users to change the Spotify recommendations. Darshna (2018) extracted the Echo Nest attributes through the Spotify API to build a music recommendation system. Many music analysts and data scientists in the industry have also used the attributes to analyze songs.[3]

The eight major Echo Nest attributes in the FMA and MSSD are explained as follows[4]:

---

[1] https://newsroom.spotify.com/company-info/
[2] http://sortyourmusic.playlistmachinery.com/
[3] See, for instance, www.theguardian.com/technology/2013/nov/25/pop-music-louder-less-acoustic, www.towardsdatascience.com/a-music-taste-analysis-using-spotify-api-and-python-e52d186db5fc and www.nycdatascience.com/blog/student-works/analyzing-spotify-song-metrics-to-visualize-popular-songs.
[444] The explanation of the Echo Nest attributes is abstracted near verbatim based on the official Spotify API https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/ and a data scientist's blog https://towardsdatascience.com/what-makes-a-song-likeable-dbfdb7abe404

- Acousticness - A measure of whether the track is acoustic. The higher the value the more acoustic the song is.
- Energy - A perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, Death Metal has high energy, while a Bach prelude scores low on the scale. The higher the value, the more energetic the song is.
- Danceability - The suitability of a track for dancing is based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. The higher the value, the easier it is to dance to this song.
- Valence - The musical positiveness conveyed by a track. The higher the value, the more positive mood for the song.
- Instrumentalness - A prediction on whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal.” The higher the instrumentalness value, the greater likelihood the track contains no vocal content.
- Speechiness -The presence of spoken words in a track. The more exclusively speech-like the recording (e.g., talk show, audio book, poetry), the higher the attribute value
- Liveness - The presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
- Tempo - The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

The values of these eight attributes, except for tempo, are reported on a 0-1 scale.

A.1.2 Librosa features. Librosa (McFee et al. 2015) is a Python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. The package can extract features from audio files and create structured data that include spectral and rhythm attributes.

A.2 Predicting Echo Nest Attributes using Librosa Features

The eight Echo Nest attributes that summarize music are very intuitive. However, we do not have the original algorithm to obtain the Echo Nest attributes for the audio in TV ads. So, we have to devise a procedure to predict the values of the Echo Nest attributes.

Since we can decompose music to 518 Librosa features, one possible solution is to exploit the correlation between Librosa features and Echo Nest attributes to build prediction models. Is it a valid approach to predict Echo Nest attributes using Librosa features? We communicated with a developer of the Echo Nest attributes asking if it would be appropriate to derive Echo Nest attributes from the Librosa features. S/he responded that: 1) a prediction model using Librosa features could work in principle; 2) it will be an approximation of the original model.

Using the soundtracks from the Free Music Archive (FMA) for which we have both the Librosa features and the Echo Nest attributes for each song, our goal is to reverse-engineer the Echo Nest algorithm to obtain the seven human-readable intuitive attributes: acousticness, danceability, energy, instrumentalness, liveness, speechiness, and valence in any piece of audio, such as the audio in a TV ad. Note that we exclude tempo as it is an objective measure that requires a different approach.

For each of the Echo Nest attributes, we evaluated eight different algorithms that fall into four major categories: (1) regularized regression (e.g., LASSO) (Tibshirani 1997), (2) ensemble trees (e.g., random forest) (Breiman 2001), (3) gradient boosting (e.g., XGBoost) (Friedman 2001; Chen and Guestrin 2016), and (4) deep learning (e.g., neural network in tensorflow) (Ripley 1996; LeCun, Bengio and Hinton 2015). We tried to be comprehensive in our evaluation and tested all of the most commonly used methods to reach the best predictive performance. We rely on cross-validation to select the best model from this set of the algorithms that are known to perform well in general. The general training pipelines are listed below:

1) Data Preprocessing Pipeline - soundtracks were first converted into the 518 Librosa features using the Python package Librosa. We withhold 20% of the data for reporting the performance of machine learning (ML) methods in Table W1 and Table W2.
2) Model Training Pipeline - the 518 spectral and rhythm features were normalized and then fed into different machine learning frameworks as input features:
    a) In R, we train models with lm (linear regression), glmnet (LASSO regression, ridge regression, elastic net), nnet (neural network), ranger (random forest), and xgboost (gradient boosting linear regression, gradient boosting trees);
    b) In Python, we use Keras API for Tensorflow to train deep learning models with different numbers of layers.
3) Hyperparameter Tuning Pipeline: we used five-fold cross-validation to conduct a grid search of hyperparameters for different ML algorithms to prevent overfitting while minimizing the RMSE. In *k*-fold cross-validation, the original data set is randomly partitioned into k equal size subsets. Among the *k* subsets, a single subset is retained as the validation data for testing the model, and the remaining *k*-1 subsets are used as training data. The cross-validation process repeats *k* times, with each of the *k* subsets used exactly once as the validation data. We obtained *k* sets of hyperparameters (e.g., the

regularized parameter λ in LASSO regression, or the depth of tree or shrinkage in gradient boosting trees), and then picked the best set of hyperparameters that minimize the RMSE in the validation data.

Finally, we picked the best tuned models out of tens of thousands of trained candidates with the objective of minimizing the root mean square error (RMSE) in the five-fold cross-validation process. We report the ML performance on the out-of-sample test data set in Table W1. From the table, we are confident in reverse-engineering 'energy'. The highest R-square value is .792 for energy followed by acousticness. However, our models have a difficult time capturing the 'liveness' of music. Using the OLS prediction as the benchmark, we obtained performance improvement in R-square by up to 13%, 5%, 4%, 34%, 62%, 63%, and 27% by using one of the machine learning methods to predict the seven attributes, respectively.

Table W2 summarizes the out-of-sample RMSE. Extreme Gradient Boosting Trees work well for most Echo Nest attributes. Deep Neural Network with three layers beat boosting methods for 'acousticness', 'speechiness', and 'valence', albeit we only used 100 epochs and the network topology was not fully tuned (e.g., different size of each layer). Using a simple OLS method as a benchmark, we see a reduction in RMSE that ranged from 3% (danceability) to 16% (speechiness). We selected XGBoost as the final predictive algorithm for 'energy' because it achieves the highest predictive power in terms of R-squared and RMSE compared to other methods in our context. XGBoost is one of the most successful prediction algorithms developed in the last decade that is widely adopted in both academia and industry. In fact, almost all the Knowledge Discovery in Database (KDD) Cup winners have used XGBoost as their learning algorithm (Nielsen 2016). The power of tree boosting comes from its adaptive determination of the local neighbourhoods that take the bias-variance tradeoff into consideration during model fitting. The correlation between the predicted and the actual Spotify audio energy value available in the FMA data set is .89 in the hold-out sample and .95 in the entire FMA data set. Figure W1 showed that the prediction is neither conservative nor extreme because most data points are aligned on the diagonal line without systematic bias.

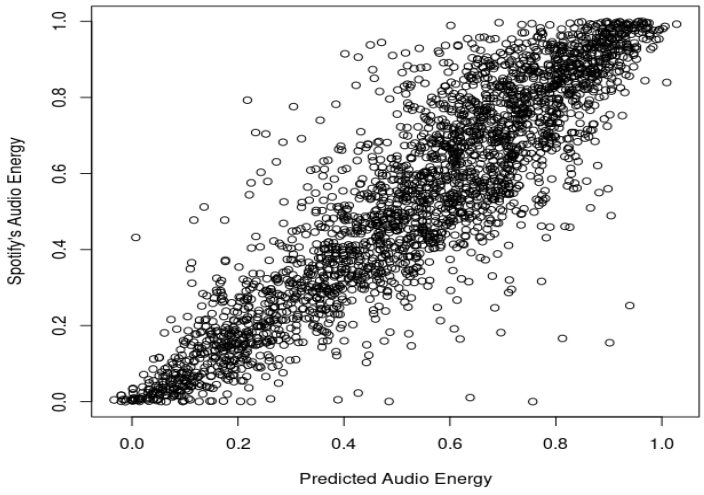
Table W1: Prediction results: out-of-sample R-squared values

	Algorithm	Acoustic-ness	Dance-ability	Energy	Instrumen-talness	Liveness	Speech-iness	Valence
[R] glmnet, nnet, ranger, xgboost	lm	.631	.503	.758	.349	.155	.321	.394
	lasso	.626	.489	.757	.378	.167	.333	.435
	ridge	.620	.479	.757	.363	.181	.338	.427
	elastic	.626	.493	.759	.378	.170	.341	.441
	nnet	.686	.517	.000	.435	.236	.002	.477
	ranger	.645	.476	.778	.388	.182	.450	.411
	xgbLinear	.651	.473	.767	.397	.183	.434	.412
	xgbTree	.670	**.530**	**.792**	**.467**	**.251**	.508	.498
[Python] Keras + Tensorflow	dnn_L1	.655	.487	.763	.436	.239	.478	.487
	dnn_L3	**.716**	.518	.787	.433	.232	**.524**	**.500**
	dnn_L5	.643	.487	.791	.425	.221	.507	.492
	dnn_L7	.679	.412	.732	.412	.216	.517	.000
	MAX	.716	.530	.792	.467	.251	.524	.500

Table W2: Prediction results: out-of-sample RMSE

	Algorithm	Acoustic-ness	Dance-ability	Energy	Instrumen-talness	Liveness	Speech-iness	Valence
[R] glmnet, nnet, ranger, xgboost	lm	.234	.134	.138	.293	.150	.113	.220
	lasso	.235	.136	.138	.285	.147	.110	.210
	ridge	.237	.137	.138	.288	.146	.110	.212
	elastic	.235	.135	.137	.285	.147	.110	.209
	nnet	.215	.132	.280	.272	.141	.135	.202
	ranger	.231	.140	.133	.290	.146	.102	.221
	xgbLinear	.227	.138	.135	.281	.146	.103	.215
	xgbTree	.221	**.130**	**.128**	**.265**	**.140**	.095	**.199**
[Python] Keras + Tensorflow	dnn_L1	.236	.139	.139	.283	.144	.101	.203
	dnn_L3	**.209**	.133	.130	.286	.146	**.095**	.200
	dnn_L5	.247	.138	.129	.291	.148	.097	.205
	dnn_L7	.224	.152	.147	.295	.149	.095	.280
	MIN	.209	.130	.128	.265	.140	.095	.199

Figure W1: Out-of-sample actual vs. predicted energy in FMA data set



A.3 Sample Ad-creatives

Below, we provide the URL links to the sample ad-creatives in Table 2.

Highest energy ad-creatives.

1. Under Armour TV Commercial, 'We Will' Featuring Michael Phelps, Misty Copeland
   https://www.ispot.tv/ad/wT0i/under-armour-we-will-featuring-michael-phelps-misty-copeland

2. Colgate TV Commercial, 'Every Drop Counts'
   https://www.ispot.tv/ad/IEcG/colgate-every-drop-counts

3. OMEGA Seamaster 300 TV Commercial, 'Spectre: Revealing the 007 Watch'
   https://www.ispot.tv/ad/AA7p/omega-seamaster-300-spectre-revealing-the-007-watch

4. Fruit of the Loom TV Commercial, 'Holidays: Feel Free to Celebrate'
   https://www.ispot.tv/ad/ZqWZ/fruit-of-the-loom-feel-free-to-celebrate

5. Ford TV Commercial, 'We Are All Champions'
   https://www.ispot.tv/ad/ATRp/ford-we-are-all-champions

6. McDonald's McCafé TV Commercial, 'Nothing Before Coffee: Downpour'
   https://www.ispot.tv/ad/wjrL/mcdonalds-mccaf-nothing-before-coffee-downpour

7. Apple Music TV Commercial, 'Apple Music Anthem' Song by Noga Erez
   https://www.ispot.tv/ad/wRjL/apple-music-apple-music-anthem-song-by-noga-erez

8. 2018 Honda Accord TV Commercial, 'Tower of Success'
   https://www.ispot.tv/ad/whtz/2018-honda-accord-tower-of-success

9. M&M's Super Bowl 2018 TV Commercial, 'Human' Featuring Danny DeVito, Todrick Hall
   https://www.ispot.tv/ad/waIJ/m-and-ms-super-bowl-2018-lucky-penny-featuring-danny-devito

10. Coca-Cola TV Commercial, 'Final Four: One Last Dance'
    https://www.ispot.tv/ad/wkIM/coca-cola-2017-final-four-one-last-dance

Lowest energy ad-creatives.

1. Nike TV Commercial, 'Until We All Win' Featuring Serena Williams
   https://www.ispot.tv/ad/wh2H/nike-until-we-all-win-featuring-serena-williams

2. Center for Biological Diversity TV Commercial, 'Polar Bear'
   https://www.ispot.tv/ad/A1Ff/center-for-biological-diversity-polar-bear

3. TaylorMade TV Commercial, 'Expect the Unexpected'
   https://www.ispot.tv/ad/AVAe/taylormade-expect-the-unexpected

4. Clorox Bleach TV Commercial, 'Bleach It Away: Distance'
   https://www.ispot.tv/ad/At94/clorox-bleach-bleach-it-away-distance

5. ASPCA TV Commercial, 'Baxter'
   https://www.ispot.tv/ad/AuZs/aspca-baxter

6. Ralph Lauren Fragrances Tender Romance TV Commercial, 'Love'
   https://www.ispot.tv/ad/AfVW/ralph-lauren-fragrances-tender-romance-love

7. Windex TV Commercial, 'The Story of Lucy: Just the Beginning'
   https://www.ispot.tv/ad/w61V/windex-the-story-of-lucy-just-the-beginning

8. PNC Bank TV Commercial, 'Know You're Saving for Special Moments'
   https://www.ispot.tv/ad/7iic/pnc-bank-know-youre-saving-for-special-moments

9. Xeljanz TV Commercial, 'Birthday Puppy'
   https://www.ispot.tv/ad/A037/xeljanz-birthday-puppy

10. Blue Apron TV Commercial, 'Farm Fresh Ingredients'
    https://www.ispot.tv/ad/dK1d/blue-apron-farm-fresh-ingredients

WA-B. UNDERSTANDING ENERGY IN ADS

We report the results from a series of analyses that are designed to provide a better understanding of the Echo Nest's audio energy measure.

B.1 Algorithm-generated vs. Human-perceived Energy in Ads

Using data collected from 2,342 participants via Amazon Mechanical Turk (MTurk), we examine whether the Echo Nest audio energy is associated with the two dimensions of affect, arousal and valence (Russell 1980).

B.1.1 Survey design. We designed a survey to measure the human-perceived energy level of 138 ads in a pilot study and in a main study. We included an attention check to make sure that participants read our instructions carefully and an equipment check to disqualify participants who can't watch video or listen to sound. In the pilot study, 28 ads were evaluated by 288 participants. In the main study, 110 ads were evaluated by 2,054 participants. The survey design is essentially unchanged between the two studies. For instance, all participants evaluated 5 ads at a time. The ads are randomly selected using stratified sampling based on the Echo Nest audio energy to ensure that we have good coverage of commercials with different levels of algorithm-generated energy. We collected data by scheduling 8 batches a day (i.e., one batch every three hours) to obtain representative samples since it has been found that arousal rises and falls in a diurnal rhythm. The main study is spread across multiple days in a week to cancel out any day of week effect.

Participants were randomly assigned to one of the three groups in which they can only hear the audio part of the ads without video (audio-only), watch the video part of the ads without audio (video-only), or encounter the intact ads with both audio and video (audio+video). We thus obtain three separate arousal factors and three valence factors for audio-only, video-only, and both audio and video given the three ad conditions. By doing so, we aim to understand how the MTurk subjects evaluate the energy of ads and learn whether Echo Nest energy is a general proxy for the overall energy of the ads.

The survey questions were borrowed from Puccinelli, Wilcox and Grewal (2015) and Belanche, Flavián and Pérez-Rueda (2017). We used multiple items with seven-point bipolar scales to assess the arousal and valence. Items 1-7 ("not energetic to energetic", "dull to exciting", "not animated to animated", "inactive to active", "relaxed to stimulated", "calm to excited", and "unaroused to aroused") were used for arousal, while items 8-12 ("unhappy to

happy", "displeasure to pleasure", "feel bad to feel good", "sadness to joy", "negative to positive") were used for valence. Some of the items were reverse-coded to prevent straight-liners and the order of the items was randomized to guard against acquiescence bias.

Table W3: Survey measures

Item	Measure	Source	Construct	Pilot Study	Main Study
0	Spotify Audio Energy	Spotify			
1	Energetic vs. Not energetic	Puccinelli et al. 2015	Arousal	✓	✓
2	Exciting vs Dull	Puccinelli et al. 2015	Arousal	✓	✓
3	Active vs Inactive	Puccinelli et al. 2015	Arousal	✓	✓
4	Animated vs Not animated	Puccinelli et al. 2015	Arousal	✓	✓
5	Stimulated vs Relaxed	Blanche et al. 2017	Arousal	✓	✓
6	Excited vs Calm	Blanche et al. 2017	Arousal	✓	✓
7	Aroused vs Unaroused	Blanche et al. 2017	Arousal	✓	✓
8	Happy vs Unhappy	Puccinelli et al. 2015	Valence	✓	✓
9	Feel good vs Feel bad	Puccinelli et al. 2015	Valence	✓	✓
10	Joy vs Sadness	Puccinelli et al. 2015	Valence	✓	✓
11	Positive vs. Negative	Puccinelli et al. 2015	Valence	✓	✓
12	Pleasure vs. Displeasure	Puccinelli et al. 2015	Valence		✓
13	Existence of human voice				✓
14-17	Repeat items 1-4 for human voice	Puccinelli et al. 2015	Arousal		✓
18	Existence of background sound				✓
19-22	Repeat items 1-4 for background sound	Puccinelli et al. 2015	Arousal		✓

Additionally, in our main study, half the participants in (1) audio-only, and (2) both audio and video conditions were asked to separately evaluate the energy level of human speech and background sound other than human speech in the ads if they can hear them. We only repeat items 1-4 for these additional evaluations because these items reach a higher reliability in the pilot study than items 5-7 and we want to limit the survey time to within 8 to 10 minutes in the main study. The set of questions help us to learn whether Spotify audio energy is more correlated with background music than human speech because the Echo Nest algorithm was mainly developed for music and sound effects. We did not disclose the values of Spotify audio energy to any participants in the survey. At the end of the survey, we asked participants to "list some

words that come to mind when you think about energetic ads" as an optional task. We expect to learn how participants evaluate the energy of ads using this text entry question after they complete the evaluation of 5 ads.

B.1.2 Survey results I: factor analysis. To obtain the human-perceived energy level of 138 ads, we ran a pilot study with 288 participants and then ran a main study with 2,054 participants.

- In the pilot study, we had 84 ad-types (28 ads x 3 versions) that were evaluated 14~19 times each. The average number of evaluations for each ad type is 17 with the 1st quartile at 16 and 3rd quartile at 18.
- In the main study, we had 330 ad-types (110 ads x 3 versions) that were evaluated 16~49 times each. The average number of evaluations for each ad-type is 30.5 with the 1st quartile at 27 and 3rd quartile at 35. In addition, we asked the participants to evaluate the energy of the ads based on human speech and background music separately.

We use the items 1-7 ("not energetic to energetic", "dull to exciting", "not animated to animated", "inactive to active", "relaxed to stimulated", "calm to excited", and "unaroused to aroused"; Cronbach's alpha > .95 for all three ad-type conditions) to measure the construct of "arousal." Items 8-12 ("unhappy to happy", "displeasure to pleasure", "feel bad to feel good", "sadness to joy", "negative to positive"; Cronbach's alpha > .95 for all three ad-type conditions) help us to measure the construct of "valence." Cronbach's alphas in the reliability analysis reveal a high internal consistency between these items separately for the two constructs. We extract the first principal component from items 1-7 as the "arousal" factor in our analysis, which accounts for more than 90% of the variance (Table W4). We also extract the first principal component of items 8-12 as the "valence" factor that explains more than 90% of the variance as well. We report the factor loadings of arousal and valence in Table W5.

Figures W2 and W3 report the results from both pilot study and the main study. The scatter plots in Figure W2 show that there is a positive correlation between Spotify audio energy with arousal in all three conditions across both studies; in contrast, the scatter plots in Figure W3 show that the correlation between Spotify audio energy and the valence is basically flat. A clear inference is that the Spotify energy measure is more reflective of arousal, which is related to how much of the ad is processed (in our context) than valence, which is related to the manner in which the ad is processed. Another inference is that the perceived arousal of the audio and the perceived arousal of the video are well correlated with each other (.57 in the main study), and the

perceived arousal of the ad audio and the perceived arousal of ad video are strongly correlated with the overall perceived arousal of the complete ad that includes both audio and video (.74 for audio and .86 for video in the main study).

Table W4: The Cronbach's alpha and explained variance in the main study

Construct	Arousal (items 1-7)		Valence (items 8-12)		Items 1-4	Items 5-7
Type/Metric	Alpha	Proportion Var.	Alpha	Proportion Var.	Alpha	Alpha
Audio-only	.9765	.9030	.9767	.9169	.9773	.9463
Video-only	.9761	.9016	.9831	.9383	.9769	.9463
Audio+Video	.9728	.8851	.9811	.9301	.9676	.9341

Table W5: Factor loadings of arousal and valence in the main study

Item	Measure	Construct	Audio-only	Video-only	Audio+Video
1	Energetic vs. Not energetic	Arousal (items 1-7)	.9677	.9694	.9613
2	Exciting vs Dull		.9569	.9592	.9543
3	Active vs Inactive		.9712	.9590	.9500
4	Animated vs Not animated		.8894	.8299	.7948
5	Stimulated vs Relaxed		.9066	.9247	.8952
6	Excited vs Calm		.9558	.9443	.9473
7	Aroused vs Unaroused		.9254	.9184	.9114
8	Happy vs Unhappy	Valence (items 8-12)	.9565	.9641	.9729
9	Feel good vs Feel bad		.9657	.9690	.9611
10	Joy vs Sadness		.9504	.9721	.9620
11	Positive vs. Negative		.9574	.9691	.9604
12	Pleasure vs. Displeasure		.9577	.9692	.9658

Figure W2: Correlations between Spotify audio energy and arousal

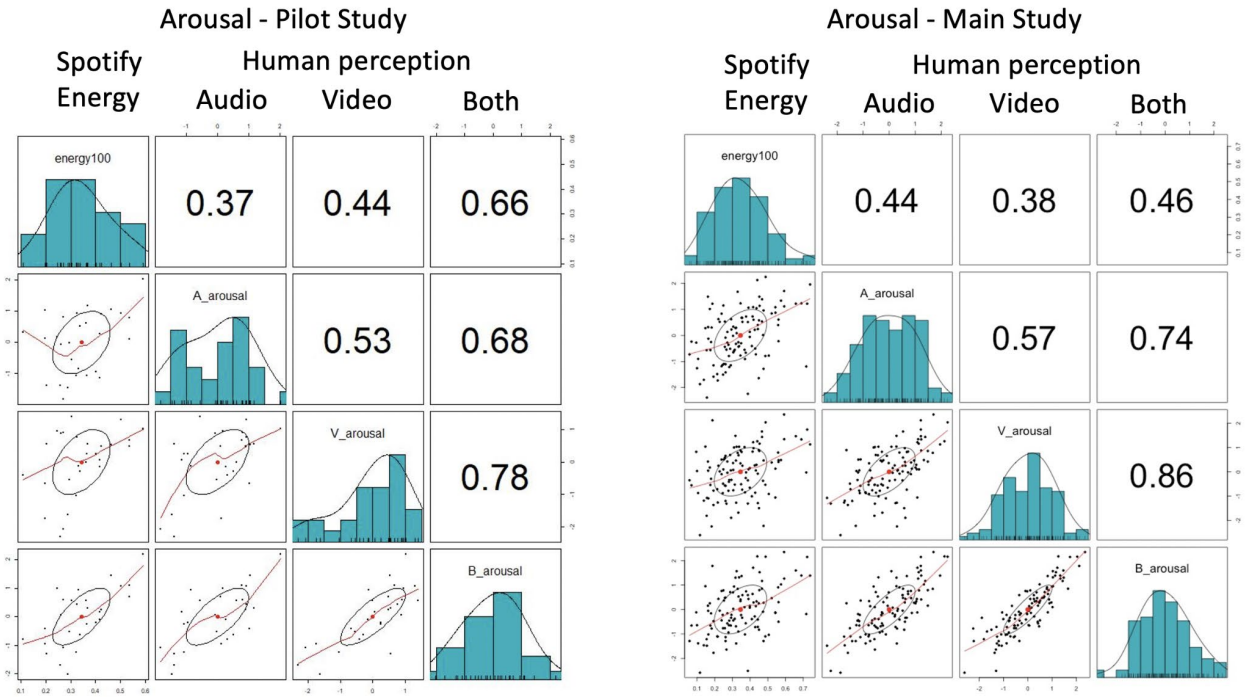
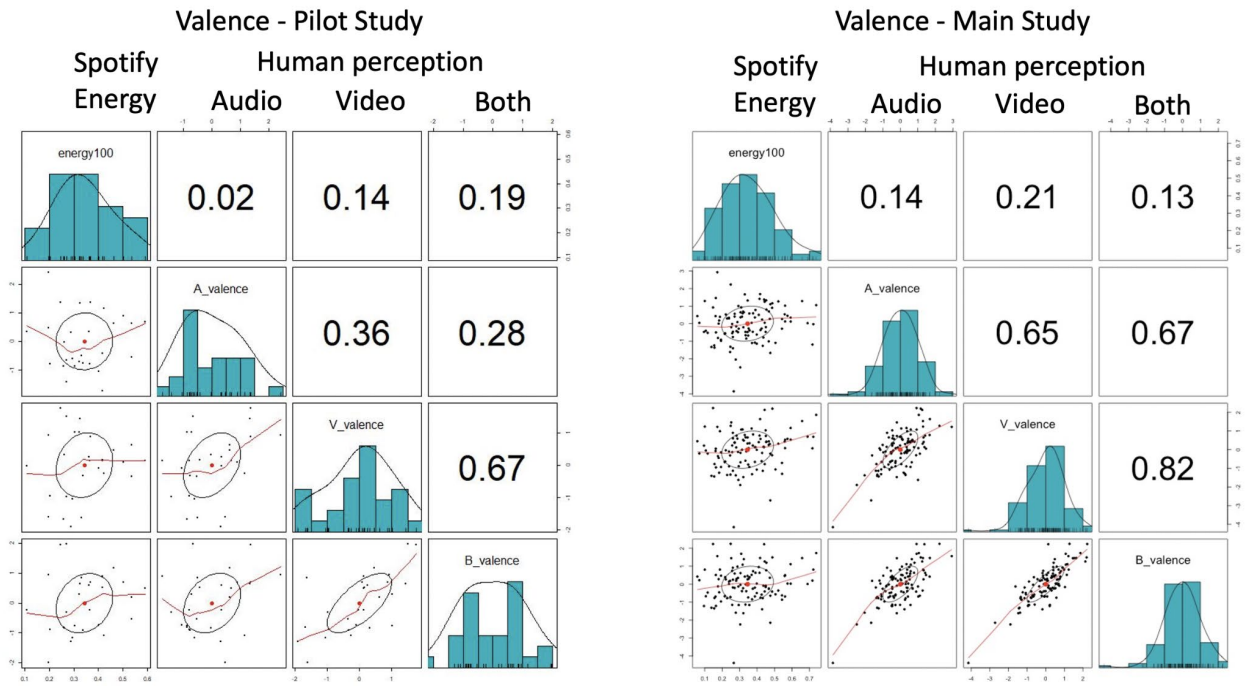


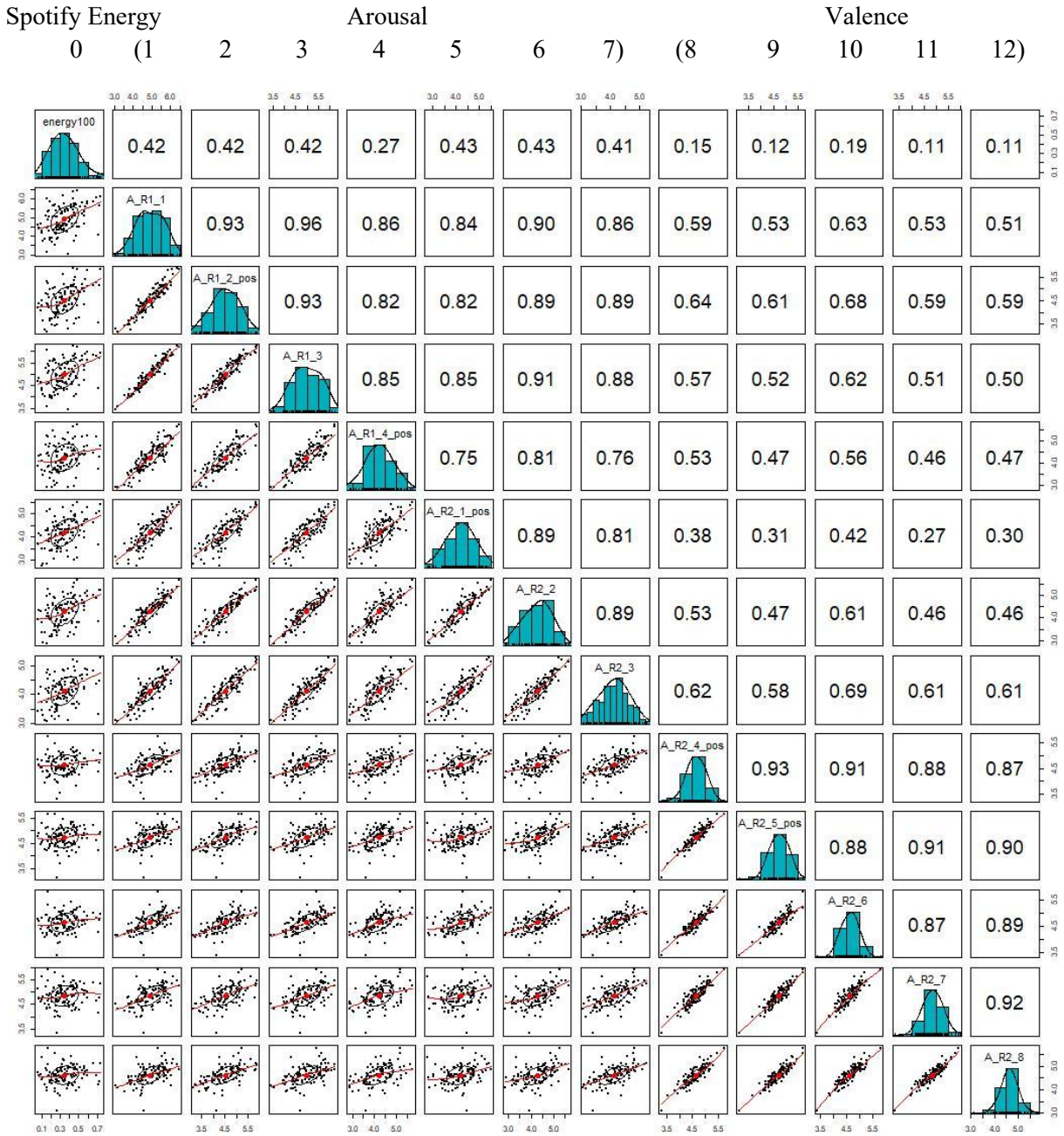
Figure W3: Correlations between Spotify audio energy and valence



B.1.3 Survey results II: item-level analysis. Besides investigating the correlation between the Spotify audio energy and the two constructs (i.e., arousal and valence), we can also evaluate its correlations with specific items in the survey to obtain more insights. Figures W4 to W6 report the correlation between Spotify energy and the 12 individual measures evaluated by survey respondents.

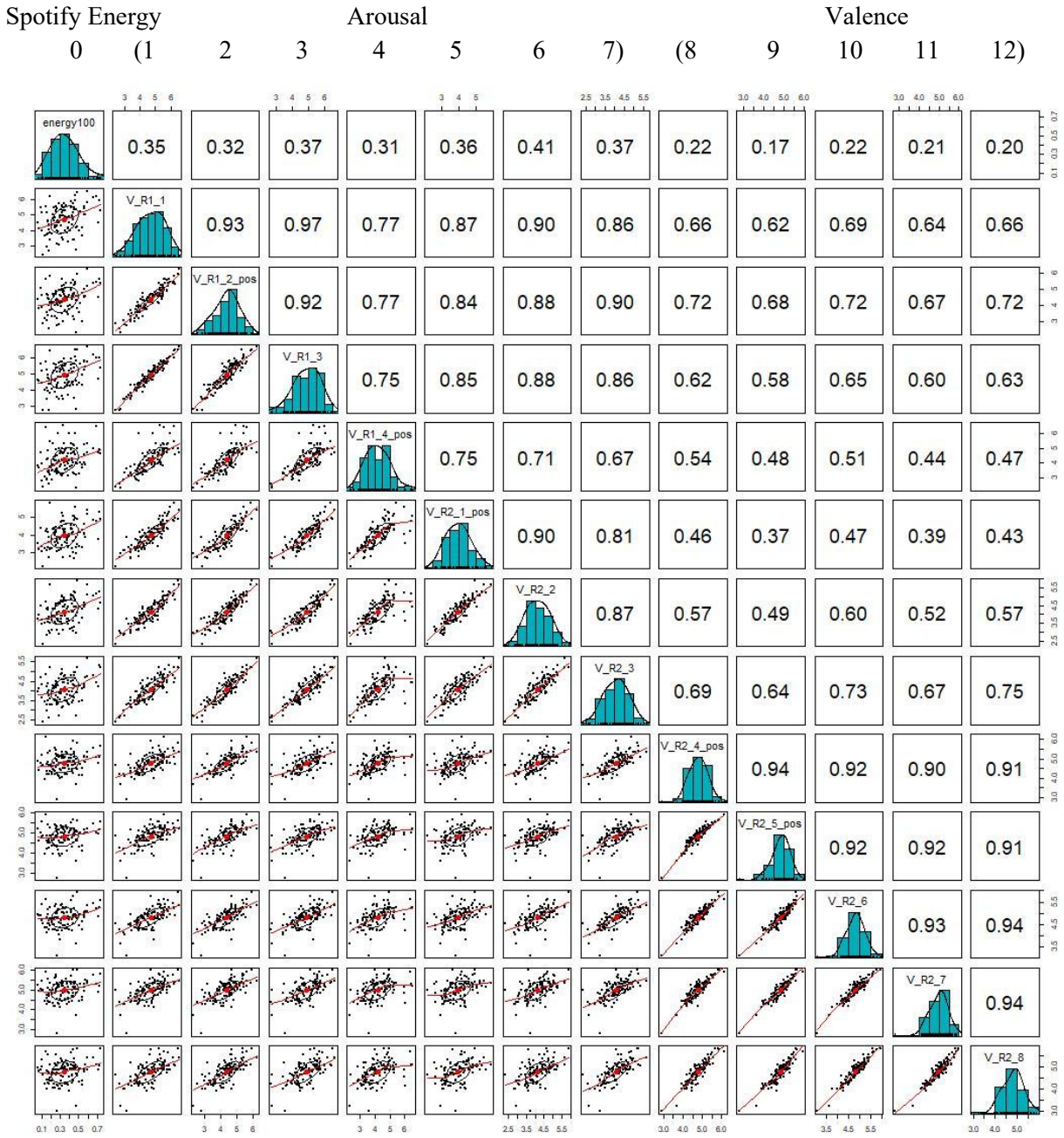
- Figure W4: As we can see from the top row, the correlations between the Spotify estimated energy measure and the seven arousal items from the *audio-only* sample is higher (average of .40) than for the five valence items (average of .136). The correlations among the individual arousal items are high, averaging .858, and among the individual valence items is .896. It is not surprising that the first principal component explains over 90% of the variance among the seven arousal items and the five valence items given these high inter-item correlations. The average correlation between the arousal items and the valence items is smaller at .527.
- Figure W5: As we can see from the top row, the correlations between the Spotify estimated energy measure and the seven arousal items from the *video-only* sample are higher (average of .356) than for the five valence items (.204). The correlations among the individual arousal items are high, averaging .841, and among the individual valence items is .923. The average correlation between the arousal items and the valence items is a lower value of .590.
- Figure W6: As we can see from the top row, the correlations between the Spotify estimated energy measure and the seven arousal items from the *audio+video* sample is higher (average of .406) than for the five valence items (average of .124). The correlations among the individual arousal items are high, averaging .813, and among the individual valence items is .912. The average correlation between the arousal items and the valence items is a lower value of .501.

Figure W4: Correlations between Spotify audio energy (item 0) and other measures (items 1-12); *Audio-only*



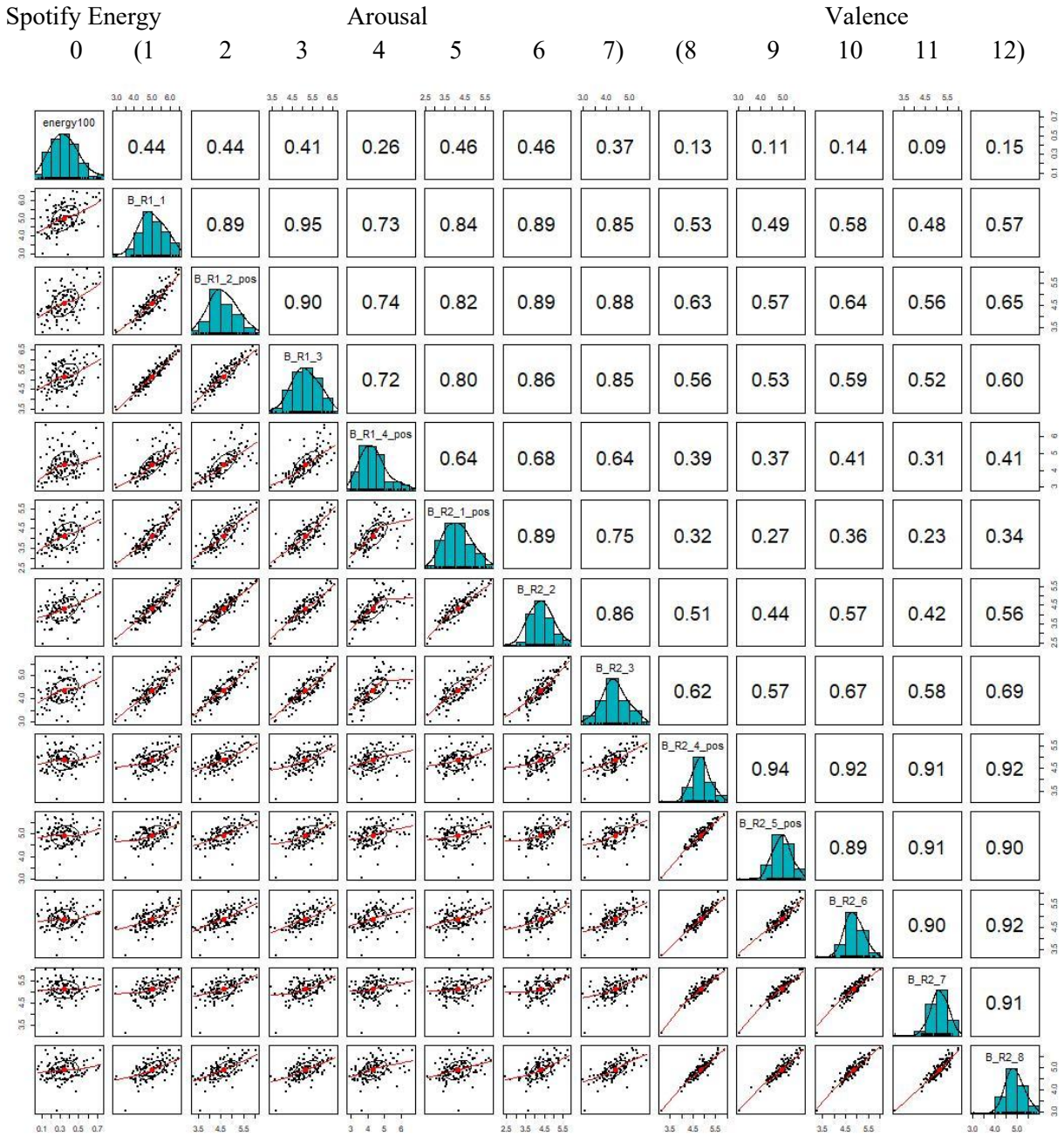
Notes: Items - 0: Spotify Audio Energy 1: Energetic vs. Not energetic 2: Exciting vs Dull 3: Active vs Inactive 4: Animated vs Not animated 5: Stimulated vs Relaxed 6: Excited vs Calm 7: Aroused vs Unaroused 8: Happy vs Unhappy 9: Feel good vs Feel bad 10: Joy vs Sadness 11: Positive vs. Negative 12: Pleasure vs. Displeasure

Figure W5: Correlations between Spotify audio energy (item 0) and other measures (items 1-12); ***Video-only***



Notes: Items - 0: Spotify Audio Energy 1: Energetic vs. Not energetic 2: Exciting vs Dull 3: Active vs Inactive 4: Animated vs Not animated 5: Stimulated vs Relaxed 6: Excited vs Calm 7: Aroused vs Unaroused 8: Happy vs Unhappy 9: Feel good vs Feel bad 10: Joy vs Sadness 11: Positive vs. Negative 12: Pleasure vs. Displeasure

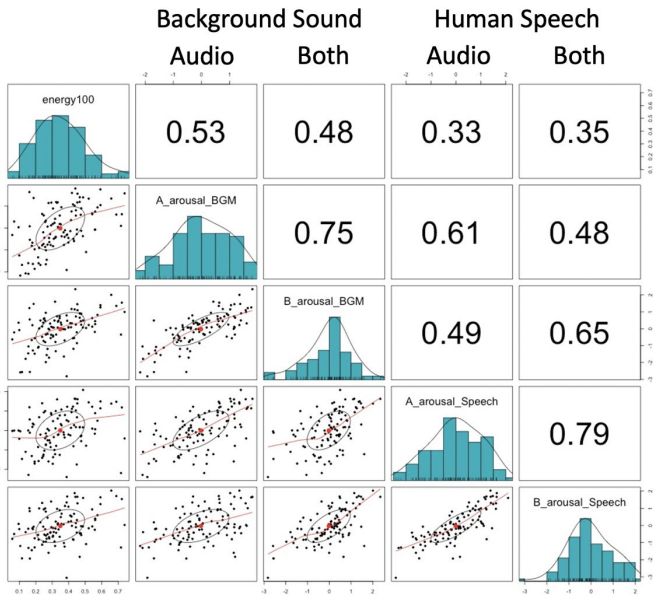
Figure W6: Correlations between Spotify audio energy (item 0) and other measures (items 1-12)
Audio+video



Notes: Items - 0: Spotify Audio Energy 1: Energetic vs. Not energetic 2: Exciting vs Dull 3: Active vs Inactive 4: Animated vs Not animated 5: Stimulated vs Relaxed 6: Excited vs Calm 7: Aroused vs Unaroused 8: Happy vs Unhappy 9: Feel good vs Feel bad 10: Joy vs Sadness 11: Positive vs. Negative 12: Pleasure vs. Displeasure

B.1.4 Survey results III: human speech vs. background sound. In the main study, we asked half the participants to do two labeling tasks in the audio-only and the audio+video conditions: whether they can hear any human speech or narration in the ads and whether they can hear any sound other than human speech (e.g., background music and/or sound effects) in the ads. Then, we created two sets of ads. The first set included ads that were labeled as having human speech. The second set included ads that were labeled as having background sound other than human speech. Next, we created the following correlation plots to display the relationships (1) between the Spotify audio energy and the arousal generated from human speech and (2) between the Spotify audio energy and the arousal generated from background sound. The correlation is higher with sounds other than human speech than human speech in both audio-only ads and the intact ads with both audio and video. The Pearson and Filon's Z test shows that the Spotify audio energy is significantly more correlated with the arousal in background sound than that in human speech in the audio-only condition (.525 vs. .326; $z = 2.616$; $p = .009$) and the intact-ad condition (.476 vs. .353; $z = 1.691$; $p = .091$). These results indicate that Spotify energy measure is reflecting the arousal from background sound effects such as music and sound from products more than due to the sound from narration.

Figure W7: Correlations between Spotify audio energy and the perceived arousal in different sources of sounds



B.1.5 Survey results IV: associated words to energetic ads. At the end of our survey, we asked participants to "list some words that come to mind when you think about energetic ads" as an optional task. 66.2% of participants in the main study (48% in pilot study) responded. The top three words mentioned are "fast", "music", and "movement." Out of 2,342 participants, the top keyword "fast" was mentioned 210 times, followed by "music" with 203 mentions and "movement" with 186 mentions. The next commonly used keywords are adjectives such as "upbeat", "exciting", "fun", and "loud." The finding shows how audiences evaluate the energy level of our ads.

Figure W8: Clouds of words reported to be associated with energetic ads

To obtain more insights, we generated word clouds for the three versions of the ads (from left to right: audio-only, video-only, and both audio and video). It is clear that music is a very important keyword when survey respondents think about energetic ads except in the video-only condition:

- Audio-only: music, upbeat, fast, loud, exciting
- Video-only: movement, fun, active, fast, exciting
- Both audio and video: fast, music, movement, upbeat, exciting

Figure W9: Clouds of words reported to be associated with energetic ads (from left to right: audio-only, video-only, and both audio and video)

It is possible that asking participants to evaluate the energy level of human speech and background sound could affect how participants think about energetic ads. To exclude the potential effect of cues on the frequency of keywords, we evaluated the response from the half of the participants who were not prompted for the evaluation of the specific auditory components in the survey. Participants who saw the intact ads with both audio and video (one sixth of the total participants), listed “movement,” “fast,” and “music” with similar frequencies.

B.2 Auditory Correlates of Audio Energy

Here we explore the relationship between the Echo Nest audio energy and two other sets of auditory characteristics. First, we report the correlation between the Echo Nest energy and other high-level Echo Nest attributes that are interpretable using the FMA data set. Second, we report the association between the Echo Nest energy and the low-level Librosa features.

B.2.1 Echo Nest Energy and Other Echo Nest Attributes. The correlation matrix for the eight Echo Nest attributes created using the 13,129 FMA tracks are shown in Table W6. The table reports that energetic audio is positively correlated with tempo (correlation = .23) and valence (.22) and negatively correlated with acousticness (correlation = -.48). The association is weak with other variables such as, liveness (.04), speechiness (-.01), and instrumentalness (-.01).

Table W6: Correlation matrix of the Echo Nest attributes in the FMA data set

	1.	2.	**3.**	4.	5.	6.	7.	8.
1. Acousticness	1.00							
2. Danceability	-.19	1.00						
3. Energy	**-.48**	**.05**	**1.00**					
4. Instrumentalness	.11	-.12	**-.01**	1.00				
5. Liveness	.05	-.14	**.04**	-.06	1.00			
6. Speechiness	.04	.17	**-.01**	-.21	.08	1.00		
7. Tempo	-.11	-.09	**.23**	.02	-.01	.03	1.00	
8. Valence	-.08	.43	**.22**	-.14	-.02	.09	.13	1.00

Notes: the number represents the Pearson correlation.

B.2.2 Echo Nest Energy and Librosa Features. We seek to understand how the Spotify energy is related to the technical components. Using the FMA data set, we (1) regress the Spotify energy measure on the 500+ librosa features using LASSO to conduct variable selection and report the most positive and negative coefficients, and (2) extract the feature importance scores from the XGBoost model that we are currently using. We report the results in Tables W7 and W8.

The variables that show up in both lists are mfcc_mean_01 and spectral_contrast_mean_02, which are commonly used auditory features in music type classification systems. The former variable is the mel-frequency cepstral coefficient that represents information related to rate of change in spectral bands and the scale provides a model of human frequency perception (Stevens, Volkmann, and Newman 1937). Mel-frequency cepstrum was originally developed for automatic speech recognition, reflects the fact that humans perceive frequency on a log scale

such that our ear differentiates much more between lower frequencies than between higher frequencies, and describes the overall spectral envelope shape of audio that captures timbre (Fong, Kumar and Sudhir 2021). The latter variable is the mean contrast of the spectral peak and the spectral valley in the audio (Jiang et al. 2002). Detailed explanations of other variables are available in the document of Librosa python package (McFee et al. 2015).

Table W7: LASSO estimation results

Librosa features	**Estimate**
mfcc_mean_01	.1577
rmse_skew_01	.0163
mfcc_max_02	.0123
mfcc_median_04	.0122
spectral_bandwidth_std_01	.0118
chroma_cqt_mean_10	.0110
rmse_median_01	.0106
spectral_contrast_std_01	.0102
chroma_cqt_mean_07	.0099
mfcc_max_01	.0094
(other 498 LibRosa variables)	…
chroma_cqt_skew_05	-.0085
chroma_cqt_skew_12	-.0104
mfcc_std_07	-.0106
chroma_cqt_skew_08	-.0108
spectral_bandwidth_mean_01	-.0110
mfcc_min_01	-.0136
mfcc_skew_01	-.0153
spectral_contrast_mean_04	-.0172
spectral_contrast_mean_02	-.0186
mfcc_kurtosis_01	-.0258

Notes: the table shows the top 20 Librosa features with the most positive and negative coefficients in a LASSO regression. Librosa features are standardized.

Table W8: Variable importance from XGBoosting tree

Librosa features	Variable importance
mfcc_mean_01	100.00
mfcc_median_01	48.34
mfcc_max_01	28.56
spectral_contrast_median_02	23.72
rmse_median_01	13.98
spectral_contrast_mean_02	10.08
spectral_rolloff_mean_01	8.07
tonnetz_std_05	6.95
tonnetz_std_06	6.48
rmse_mean_01	3.85
mfcc_skew_01	3.71
tonnetz_std_04	2.63
spectral_contrast_median_04	2.37
zcr_median_01	2.29
chroma_cqt_median_08	2.23
mfcc_median_03	2.21
chroma_cqt_median_07	2.21
spectral_centroid_median_01	1.76
chroma_cqt_mean_04	1.66
chroma_cqt_median_06	1.47

Notes: the table reports the top 20 Librosa features with the highest variable importance from a XGBoosting tree model.

B.3 Visual Correlates of Audio Energy

Here, we explore what visual elements of ads constitute energetic vs. not energetic ads. In doing so, we leverage two data sources for visual elements of ads: iSpot.tv metadata and our own visual feature extraction.

B.3.1 Echo Nest Audio Energy and iSpot.tv Metadata. As reported below (also in Table 2 of the paper), the iSpot.tv data include metadata that characterize some aspects of ad content.

- Promotion: an indicator for whether the ad-creative includes a sales promotional message
- Animal: an indicator for whether the ad-creative includes a sales promotion message
- Song: an indicator for whether the ad-creative has an accompanying popular song
- Mood:
    - Active: an indicator for whether the ad-creative is action-oriented
    - Emotional: an indicator for whether the ad-creative is emotional
    - Informational: an indicator for whether the ad-creative is informational
    - Funny: an indicator for whether the ad-creative is humorous
    - Sexy: an indicator for whether the ad-creative has a sexual theme

The variables are coded by a unit within iSpot.tv named the Content Operations Team that views every ad and adds the metadata. One advantage of using the data is that the variables are available for most ad creatives in the data set used for estimation. The correlation matrix for audio energy and the iSpot.tv variables are reported in Table W9. The table reports that energetic audio is positively correlated with song (.101), promotion (.077) and active mood (.071), whereas the correlation is negative with emotional mood (-.150) and funny mood (-.096).

Table W9: Correlation matrix of the Echo Nest audio energy and iSpot.tv metadata

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Energy	**1.000**								
2. Promotion	**.077**	1.000							
3. Animal	**-.002**	-.019	1.000						
4. Song	**.101**	.004	.015	1.000					
5. Mood-active	**.071**	.056	-.037	.065	1.000				
6. Mood-emotional	**-.150**	-.057	.004	.037	-.269	1.000			
7. Mood-informational	**.041**	-.055	.081	-.046	-.754	-.086	1.000		
8. Mood-funny	**-.096**	.037	-.058	-.091	-.400	-.046	-.128	1.000	
9. Mood-sexy	**<.001**	-.018	-.019	.057	-.128	-.015	-.041	-.022	1.000

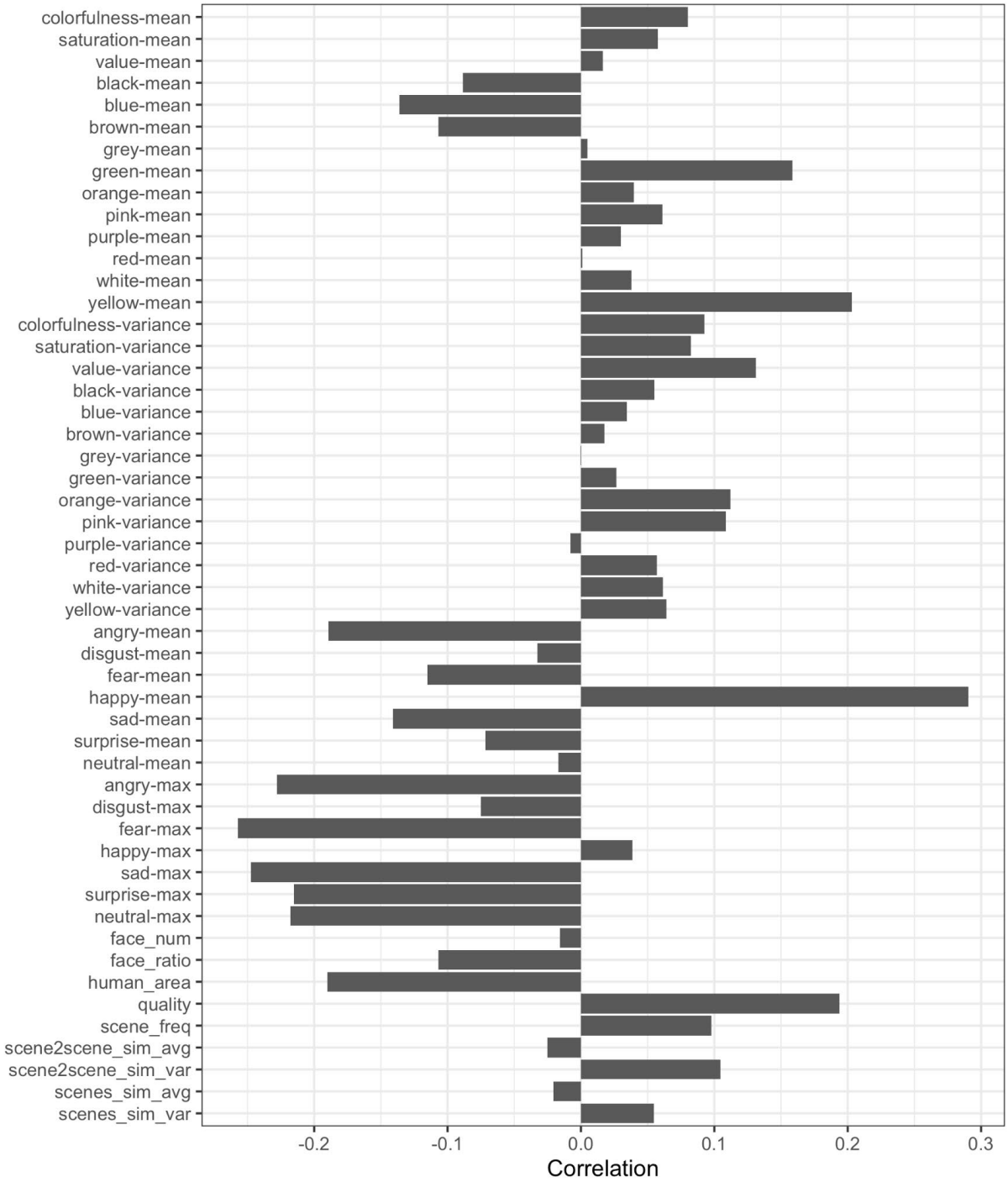
B.3.2 Echo Nest Audio Energy and Algorithm-generated Visual Features. We extracted a set of visual features in ads using a collection of open-source algorithms collated by Schwenzow et al. (2021). These features are broadly categorized into colors, emotions, faces, quality and scene. In Table W10, we list the video features extracted and their means and standard deviations for the 110 commercials we used for the survey in Section B.3.1. In Figure W10, we report the correlations between video features and the Echo Nest audio energy. As shown, the correlations vary across different video features: some video features are positively correlated with audio energy measured with the Echo Nest algorithm, whereas others are negatively correlated. For instance, happy emotions are the most positively correlated, followed by the mean level of yellow color and video quality. The colorfulness (both mean and variance) and the variance in the similarity between scenes are also positively correlated. On the other hand, emotions other than ‘happy’ (e.g., ‘fear,’ ‘sad’ or ‘angry’ emotions) are negatively correlated with audio energy.

Table W10: List of extracted visual features

	Category	Features	Aggregation	Mean	SD	Aggregation (2nd)	Mean	SD
1	Colors	Colorfulness	Mean	.359	.149	Variance	.025	.030
2	Colors	Saturation	Mean	.514	.170	Variance	.020	.020
3	Colors	Value	Mean	.369	.117	Variance	.022	.014
4	Colors	Black	Mean	.230	.207	Variance	.026	.032
5	Colors	Blue	Mean	.117	.133	Variance	.022	.035
6	Colors	Brown	Mean	.130	.118	Variance	.014	.018
7	Colors	Grey	Mean	.125	.093	Variance	.012	.014
8	Colors	Green	Mean	.086	.145	Variance	.012	.026
9	Colors	Orange	Mean	.027	.089	Variance	.003	.011
10	Colors	Pink	Mean	.045	.102	Variance	.005	.020
11	Colors	Purple	Mean	.025	.053	Variance	.005	.024
12	Colors	Red	Mean	.041	.094	Variance	.007	.022
13	Colors	White	Mean	.154	.168	Variance	.031	.043
14	Colors	Yellow	Mean	.020	.038	Variance	.002	.005
15	Emotions	Angry	Mean	.082	.094	Max	.469	.387
16	Emotions	Disgust	Mean	.004	.010	Max	.070	.155
17	Emotions	Fear	Mean	.194	.164	Max	.679	.345
18	Emotions	Happy	Mean	.263	.217	Max	.750	.376
19	Emotions	Sad	Mean	.190	.144	Max	.693	.319
20	Emotions	Surprise	Mean	.025	.038	Max	.279	.349
21	Emotions	Neutral	Mean	.243	.178	Max	.791	.320
22	Faces	Face_num		1.447	.794			
23	Faces	Face_ratio		.351	.237			
24	Faces	Human_area		.266	.183			
25	Quality	Quality		221.013	158.366			
26	Scene	Scene_freq		.476	.229			
27	Scene	Scene2Scene_Similarity	Mean	.727	.024	Variance	.004	.003
28	Scene	Scenes_Similarity	Mean	.760	.031	Variance	.062	.021

Notes: see Schwenzow et al. (2021) for the definition of these video features.

Figure W10: Visual correlates of energy -
Correlation between video features and the Echo Nest energy measure



WA-C. FIRST-STAGE ESTIMATION RESULTS

This section reports the first stage estimation results for our between-estimator analysis.

- Table W11 reports the portion of variation in ATR_{it} explained by a given set of fixed effects we use in the first stage. Overall, we find ad-creative fixed effects alone explain the largest portion of the variation at 55%, whereas all of the fixed effects combined explain 57% of the variation.
- Table W12 reports the portion of variation in ATR_{it} explained by a given set of variables we use in the second stage. Among the variables, we find that brand fixed effects explain the largest portion of variation in ATR_{it}.
- Figures W11 to W14 plot the estimated fixed effects for networks and programs (Figure W11), year-week (Figure W12), day of week and dayparts (Figure W13), and in-program positions (Figure W14).

Table W11: Variation in ad-tuning rate explained by fixed effects in the first-stage

	Ad-creative fixed effects	Network fixed effects	Program fixed effects	Time† fixed effects	Position fixed effects	Explained Variation
1	O					**.5496**
2		O				.0017
3			O			.0218
4				O		.0259
5					O	.0015
6	O	O				.5500
7	O		O			.5589
8	O			O		.5614
9	O				O	.5498
10		O	O			.0220
11		O		O		.0273
12		O			O	.0029
13			O	O		.0445
14			O		O	.0222
15				O	O	.0271
16	O	O	O			.5590
17	O	O		O		.5618
18	O	O			O	.5503
19	O		O	O		.5616
20	O		O		O	.5589
21	O			O	O	.5616
22		O	O	O		.0447
23		O	O		O	.0225
24		O		O	O	.0284
25			O	O	O	.0450
26	O	O	O	O		.5695
27	O	O	O		O	.5590
28	O	O		O	O	.5621
29	O		O	O	O	.5695
30		O	O	O	O	.0451
31	O	O	O	O	O	**.5696**

Notes: The table reports the variation in the ad-tuning rate (measured with the 75% cutoff) explained by a given set of fixed effects we used in the first stage estimation.

†Time fixed effects include year-week, day of the week, and day part fixed effects.

Table W12: Variation in ad-tuning rate explained by fixed effects in the second-stage

	Brand fixed effects	Duration fixed effects	Content† fixed effects	Mood‡ fixed effects	Explained Variation
1	O				**.1661**
2		O			.0099
3			O		.0117
4				O	.0008
5	O	O			.1700
6	O		O		.1674
7	O			O	.1661
8		O	O		.0190
9		O		O	.0102
10			O	O	.0123
11	O	O	O		.1713
12	O	O		O	.1700
13	O		O	O	.1675
14		O	O	O	.0192
15	O	O	O	O	**.1713**

Notes: The table reports the variation in the ad-tuning rate (measured with the 75% cutoff) explained by a given set of fixed effects we used in the second-stage estimation.
†Content fixed effects include promotion, animal and song fixed effects.
‡Mood fixed effects include active, emotional, funny, informational and sexy fixed effects.

Figure W11: First-stage estimation results - networks and programs

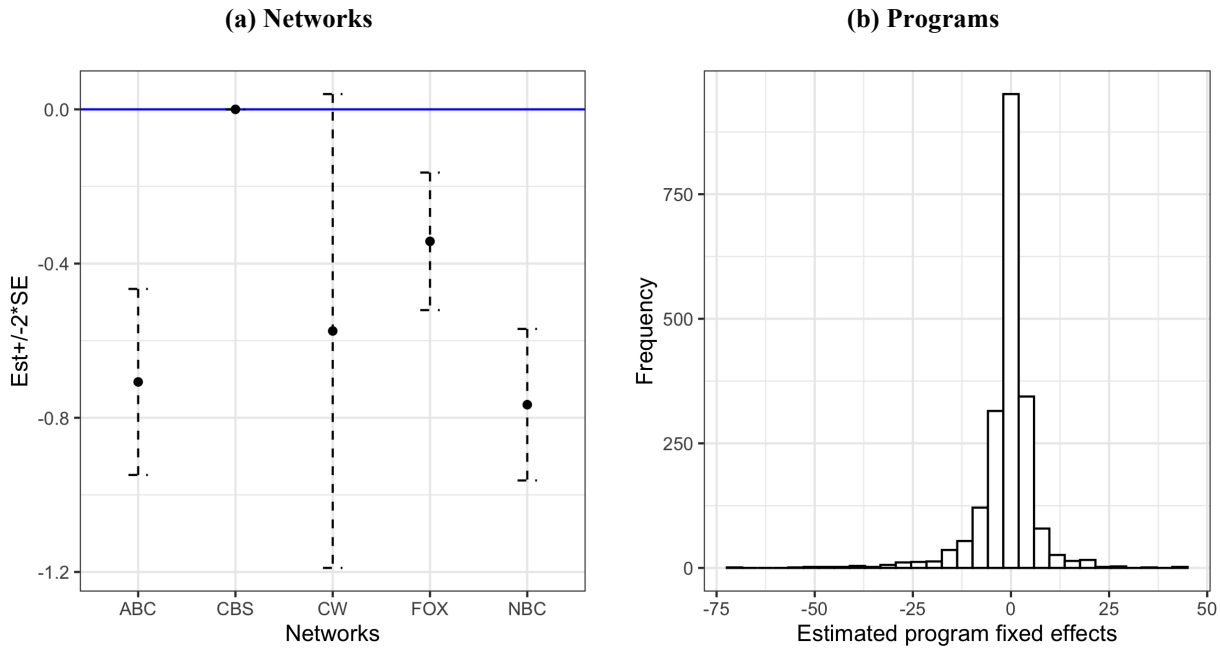


Figure W12: First-stage estimation results – year-week

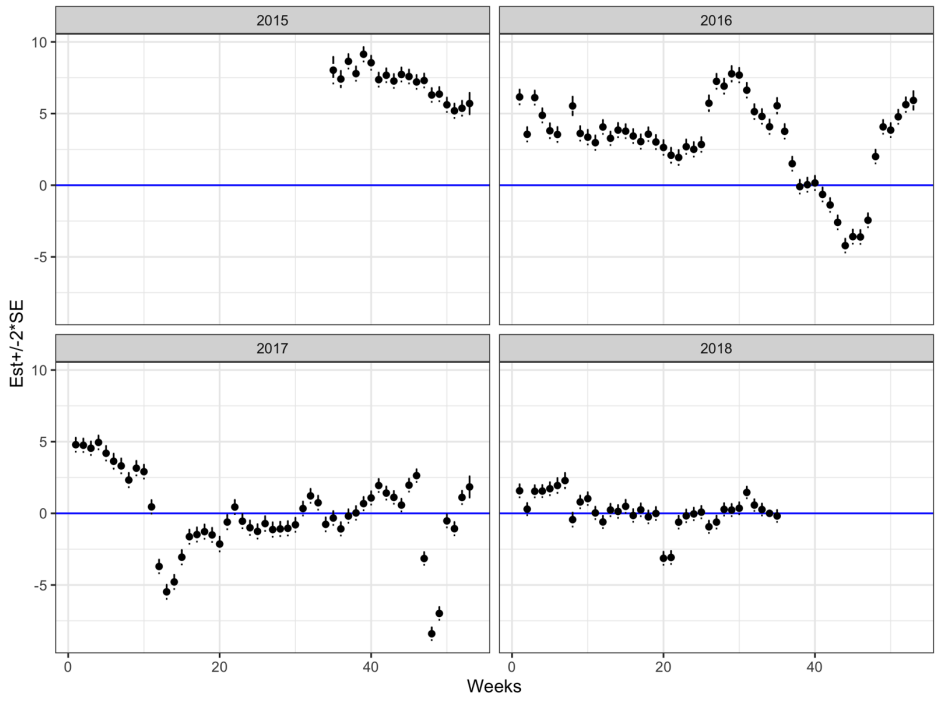
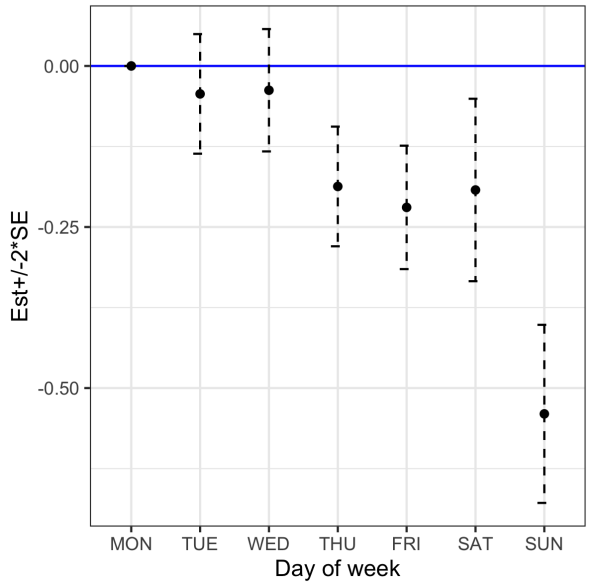


Figure W13: First-stage estimation results - day of week and dayparts

(a) Day of week



(b) Dayparts

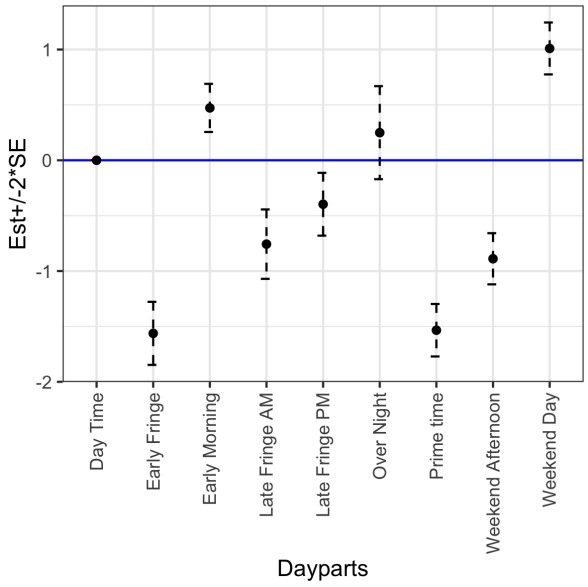
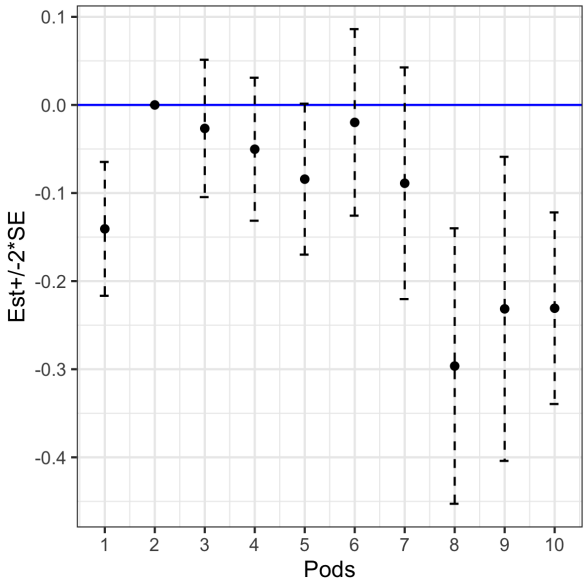
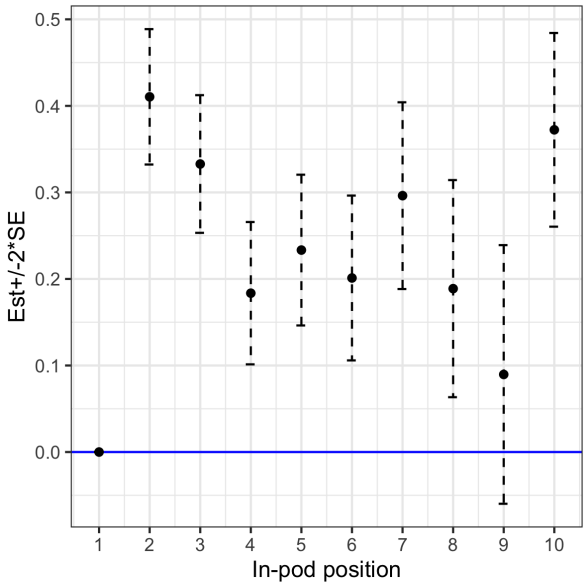


Figure W14: First-stage estimation results - in-program positions

(a) Pod



(b) Pod order



WA-D. DEPENDENT VARIABLE MEASURED WITH ERROR

This section discusses the heteroskedasticity issue in the second stage of our between-estimator analysis and explains a potential correction for the problem, following Lewis and Linzer (2005).

D.1 The Problem

In the second stage of our between-estimator, we want to estimate the following equation:

$$\delta_i = \alpha + \beta \cdot Energy_i + Z_i'\eta + \varepsilon_i, \quad \text{(W1)}$$

where δ_i represents the ad-creative fixed effect for ad i, $Energy_i$ is its energy level, Z_i is a set of control variables, and $\varepsilon_i \sim N(0, \sigma^2)$ is an error term. Note that δ_i is unobservable, so we obtain an unbiased estimate $\hat{\delta}_i$ from the first stage regression, where:

$$\hat{\delta}_i = \delta_i + u_i \quad \text{(W2)}$$

and $E(u_i) = 0$ and $Var(u_i) = \omega_i^2$. Plugging (W2) into (W1) yields,

$$\hat{\delta}_i = \alpha + \beta \cdot Energy_i + Z_i'\eta + v_i, \quad \text{(W3)}$$

where $v_i = u_i + \varepsilon_i$.

In estimating Equation W3, a heteroskedasticity issue arises when $\omega_i \neq \omega_j$ for some i and j. Correcting for this requires both σ^2 and ω_i^2 to be known. A weighted least squares (WLS) approach can yield best linear unbiased estimators, where the weights are given as follows:

$$w_i = 1/\sqrt{\omega_i^2 + \sigma^2}, \quad \text{(W4)}$$

The problem is that as econometricians we do not know ω_i^2 and/or σ^2. Below, we discuss potential solutions to the problem.

An OLS approach can only be used when one believes either $\omega_i^2 = 0$ for i or it does not vary across units. Then, the weights simply become $w_i = 1/\sigma$. In our context, this simplification is unlikely to hold since we rely on a finite sample to estimate δ_i in our first stage, and the sample size varies across units. Therefore, we can reasonably assume that ω_i^2 is nonzero, at least for some units, and its magnitude is not identical across units.

D.2 A Solution: a Feasible Generalized Least Squares (FGLS) Approach

A WLS approach sets $w_i = 1/\omega_i$ with an underlying assumption that σ^2 is close to zero or relatively smaller than ω_i^2. Typically, ω_i^2 is estimated from the first stage regression as the standard deviation of u_i in Equation W2. This method is perhaps the most widely used as pointed out by Lewis and Linzer (2005), but the authors also warn that the assumption can be quite

costly, since the WLS estimator becomes increasingly inefficient as the relative size of σ^2 increases. The method proposed by Lewis and Linzer (2005) is meant to address the concern regarding WLS. To be specific, if ω_i^2 is known then one can estimate σ^2, and use it in the weights together with ω_i^2, as shown in Equations W5 and W6:

$$w_i = 1/\sqrt{\omega_i^2 + \hat{\sigma}^2}, \tag{W5}$$

where:

$$\hat{\sigma}^2 = [\Sigma_i \hat{v}_i{}^2 - \Sigma_i \omega_i^2 + tr((X'X)^{-1}X'GX)]/(N-k). \tag{W6}$$

The terms in Equation W6 are as follows: $\hat{v}_i$ is the OLS regression residual of Equation W3, X is the $N \times k$ matrix of regressors in Equation W3, G is a $N \times N$ diagonal matrix with ω_i^2 as the ith diagonal element, N is the number of units and k is the number of parameters to estimate.

Under the maintained assumptions, this FGLS (feasible generalized least squares) approach yields an unbiased, asymptotically efficient estimator for Equation W1. Obviously, however, this approach is not perfect since we do not know ω_i^2. As commonly done for the WLS approach, we use the estimated standard error of the δ_i's from the first stage to obtain $\widehat{\omega}_i$, and plug it into Equation W5. In doing so, we restrict our attention to the ad-creatives for which we observe at least two insertions during our observation period with complete data record on the regressors and the regressand in both stages. For those with only one insertion, computing the standard errors is infeasible.

Due to the large number of ad-creative fixed effects, we bootstrap the standard errors based on 5,000 samples. Not surprisingly, ad-creatives that appear only a few times in our data tend to have larger standard errors, which is directly taken care of by the second stage estimation as the weights.

D.3 Estimation Results

Tables W13 and W14 report the second stage estimation results using either OLS or the FGLS approaches. The association between audio energy and the ad-creative fixed effects is consistently positive and statistically significant across various specifications with both approaches. In the main text, we report the results from the FGLS method. The results with the correct FGLS approach turn out to be similar to those from the incorrect OLS approach.

Table W13: Second-stage estimation results using OLS

	DV: estimated ad-creative fixed effects ($\hat{\delta}_i$)			
	(1)	(2)	(3)	(4)
Energy	.890	.727	.848	3.629***
	(.839)	(.834)	(.843)	(.959)
Duration: 30 sec		-3.114***	-3.198***	-2.086***
		(.237)	(.237)	(.289)
Duration: 60 sec		-7.625***	-7.227***	-4.448***
		(.521)	(.534)	(.768)
Duration: 90 sec		-6.931***	-5.351***	-11.892***
		(1.732)	(1.750)	(2.748)
Duration: > 90 sec		-15.310***	-14.500***	-15.661***
		(1.465)	(1.477)	(2.037)
Promotion		-5.011***	-4.930***	-2.204***
		(.261)	(.262)	(.375)
Animal		1.636***	1.418***	.528
		(.405)	(.406)	(.492)
Song		-.971***	-1.134***	-.681*
		(.317)	(.319)	(.375)
Mood: Emotional			1.971***	.872
			(.688)	(.812)
Mood: Informational			-2.166***	-1.238**
			(.491)	(.596)
Mood: Funny			.775***	.832**
			(.294)	(.356)
Mood: Sexy			-.338	-1.717
			(1.372)	(1.968)
Constant	83.038***	86.462***	86.409***	
	(.293)	(.325)	(.337)	
Brand FE	No	No	No	Yes
N	18,933	18,923	18,861	18,861
R^2	<.01	.043	.046	.275
Adj. R^2	<.01	.043	.045	.172

Notes: The table reports the estimation results of Equation W3 using OLS; energy is measured using the first 75% of ads' duration; standard errors are reported in parentheses; *$p<.10$, **$p<.05$, ***$p<.01$.

Table W14: Second-stage estimation results using FGLS

	DV: estimated ad-creative fixed effects ($\hat{\delta}_i$)			
	(1)	(2)	(3)	(4)
Energy	.928	1.094	1.180	3.343***
	(.811)	(.802)	(.811)	(.922)
Duration: 30 sec		-2.790***	-2.877***	-1.806***
		(.228)	(.229)	(.278)
Duration: 60 sec		-6.308***	-5.939***	-3.796***
		(.546)	(.560)	(.789)
Duration: 90 sec		-3.872**	-2.378	-10.680***
		(1.657)	(1.674)	(2.710)
Duration: > 90 sec		-21.842***	-20.998***	-21.305***
		(1.774)	(1.789)	(2.317)
Promotion		-5.033***	-4.954***	-2.401***
		(.258)	(.259)	(.369)
Animal		1.841***	1.678***	.940**
		(.389)	(.390)	(.479)
Song		-1.018***	-1.151***	-.647*
		(.298)	(.300)	(.357)
Mood: Emotional			1.427**	.465
			(.673)	(.791)
Mood: Informational			-2.105***	-1.056*
			(.518)	(.615)
Mood: Funny			.589**	.520
			(.278)	(.342)
Mood: Sexy			-.271	-1.848
			(1.279)	(1.850)
Constant	83.269***	86.181***	86.156***	
	(.284)	(.311)	(.322)	
Brand FE	No	No	No	Yes
N	18,933	18,923	18,861	18,861
R^2	<.01	.042	.044	.257
Adj. R^2	<.01	.041	.043	.152

Notes: The table reports the estimation results of Equation W3 using FGLS; energy is measured using the first 75% of ads' duration; standard errors are reported in parentheses; * *p<.10,* ** *p<.05,* *** *p<.01.*

WA-E. ADDITIONAL ANALYSES

E.1. Alternative Baselines for the Within-estimator Approach

We report the results of our within-estimator approach using two sets of alternative baselines. First, we compute the baseline as the average energy level of ad-creatives in a particular (1) day of week and hour of the day, (2) month, day of week and hour of the day, and (3) network, month, day of week and hour of the day. Second, we use the energy level of the previous ad within the same commercial break as the baseline. So, we consider four different baselines. In each case, we compute the difference in the energy levels between an ad-creative and the baseline as $\Delta Energy_{it} = Energy_i - Baseline_{it}$. Note that we used the first 75% of ads' length when computing $Energy_i$, whereas ad-creatives' entire duration is used when computing $Baseline_{it}$. This is because our dependent variable is whether an ad was interrupted before it reaches 75% of its length.

We estimate the following regression:

$$ATR_{it} = \delta_i + X_t'\gamma + \beta \cdot I[\Delta Energy_{it} > 0] + \varepsilon_{it}, \quad \text{(W7)}$$

where δ_i are the ad-creative fixed effects and X_t includes network, program, year-week, day-of-week, daypart, and position within program fixed effects, and I is an indicator function. The term β is the parameter of interest.

Table W15 reports the estimation results. When we use the first three baselines noted above which reflect the mean energy level of TV ads over different intervals as the baseline, we find a higher energy level than the baseline is associated with a greater ad-tuning rate. With the energy level of the previous ad within the same commercial break as baseline, we find the estimate to be statistically insignificant.

Table W15: Within-estimator approach - estimation results using alternative baselines

	DV: Ad-tuning rate			
	(1)	(2)	(3)	(4)
Baseline	Mean energy level of TV ads by day of week x hour of the day	Mean energy level of TV ads by day of week x hour of the day x month	Mean energy level of TV ads by day of week x hour of the day x month x network	Energy level of the preceding ad
β: $1[\Delta Energy_{it}>0]$	.275***	.181**	.192***	–.041
	(.085)	(.074)	(.065)	(.040)
Ad-creative FE	Yes	Yes	Yes	Yes
Other FE†	Yes	Yes	Yes	Yes
N	1,057,798	1,057,798	1,057,798	468,514
R^2	.638	.638	.638	.660
Adj. R^2	.631	.631	.631	.645

Notes: The table reports the estimation results of Equation W7 with various baselines for computing the deviation in the energy level of each insertion of an ad from the baseline; standard errors are reported in parentheses.

** p<.10, ** p<.05, *** p<.01.*

†Other fixed effects include network fixed effects, program fixed effects, year-week fixed effects, day of the week fixed effects, daypart fixed effects, position fixed effects.

E.2. Alternative Definitions of Ad-tuning Rate

In this section, we report the second-stage estimation results of our between-estimator analysis using alternative definitions of ad-tuning rate. First, we report the estimation results using the 50% and 25% cutoff values in addition to the 75%. Second, we report the results with transformed ATRs to alleviate the concern that ATRs being truncated between 0 and 100 can violate the normality assumption of regression analyses. We begin with reporting the distribution of ATRs measured with different cutoff values.

As shown in Table W16, the average ATR is smaller when measured with the 75% cutoff than with 50% or 25%, and that of 50% is smaller than 25%. This is to be expected because viewers who were tuned in for an ad for more than 75% should have tuned in for 50% and 25% as well. The variation of each ATR measure, however, varies substantially. The standard deviation of ATR (75%) is about 18.8, whereas that of ATR (25%) is just about 3.5. Indeed, over 60% observations of ATR (25%) in our data are 100, the maximum value of ATR.

Figure W15 visualize the distributions of ATRs in three different ways: raw data, with arcsine transformation (i.e., arcsine(sqrt(ATR/100))) and with logit transformation (i.e., log(ATR/(100-ATR))). As shown, the logit transformation does a better job than the arcsine transformation in transforming ATRs into a more bell-shaped distribution. One drawback of logit transformation, however, is that it drops observations with ATRs of zero and 100, which drops more than 61% of observations for the 25% cutoff (Table W16).

In Tables W17-W19, we report the second-stage estimation results of Equation 3 in the main text with the ad-tuning rate in the first-stage measured at the 50% and 25% cutoffs, either with or without transforming ATRs. The energy levels of ad-creatives are measured with the first 75%, 50% and 25% of ads' length depending on the cutoff defining ATRs. We used the FGLS method to account for the measurement errors in the dependent variable. Focusing on our preferred specification (column (4)), we find the associations between audio energy and ad-creative fixed effects are positive and statistically different from zero for the 75% cutoff value regardless of transformation of ATRs. However, the estimates are not statistically different from zero for both 50% and 25%. The findings are consistent no matter how we transform our dependent variable (ATRs).[5] Although we are unable to explain the different results between the 75% cutoff and 50%/25% cutoffs, we speculate that more of an ad has to be seen for the stimulation in the ad to have an effect, so the shorter the tuning time the less the effect.

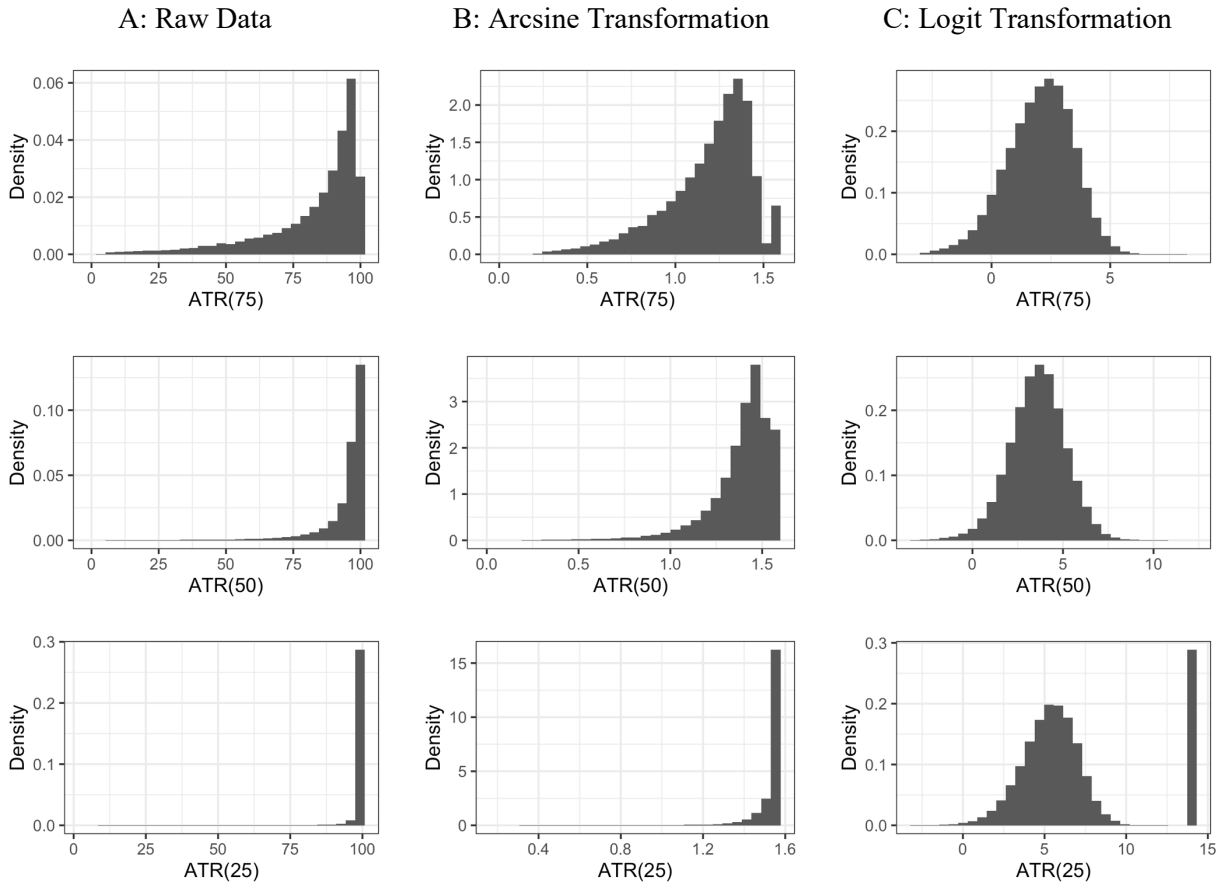
Lastly, in Table W20, we vary the cutoff values depending on ads' length. For 15-sec or shorter ads, we used a 75% cutoff. For (15,30]-sec ads, we used 50%. For ads longer than 30 seconds, we used 25%. It allows us to keep the duration of initial exposure to ads before taking actions of avoidance roughly comparable. We make sure that ads' energy levels are also measured by taking into account the varying cutoff values (e.g., the first 75% (50%) of ads to measure the energy level of a 15(30)-second ad). We find the results are qualitatively unchanged.

[5] For the 25% cutoff, we also dropped observations with ATR(25)=99.9999, which is shown as the spike at logit(ATR)=13.816 to alleviate the violation of normality assumption.

Table W16: Distribution of ATR with different cutoff values

	N	Mean	Std.Dev.	Min.	Max.	Obs. w/ ATR=0	Obs. w/ ATR=100
ATR (75%)	1,085,474	82.67	18.80	0	100	4 (<.001%)	38,215 (3.5%)
ATR (50%)	1,085,474	94.34	10.48	0	100	2 (<.001%)	133,225 (12.3%)
ATR (25%)	1,085,474	99.34	3.53	4.17	100	0 (0%)	665,948 (61.4%)

Figure W15: Distribution of ATRs without and with transformation



Notes: The figure reports the distributions of ad-tuning rate (ATR) across three different cutoff values (75%, 50% and 25%) without or with transformation. The observed spike in the right side of ATR(25) with logit transformation is a mass of observations with ATR(25)=99.9999. We drop the observations from the corresponding regression.

Table W17: Second-stage estimation results for different cutoff values

	DV: estimated ad-creative fixed effects ($\hat{\delta}_i$)		
	(1)	(2)	(3)
	75%	50%	25%
Energy	3.343***	.311	-.078
	(.922)	(.490)	(.153)
Duration: 30 sec	-1.806***	-4.248***	-.786***
	(.278)	(.151)	(.048)
Duration: 60 sec	-3.796***	-9.463***	-4.532***
	(.789)	(.428)	(.136)
Duration: 90 sec	-10.680***	-16.868***	-10.397***
	(2.710)	(1.470)	(.467)
Duration: > 90 sec	-21.305***	-27.530***	-15.037***
	(2.317)	(1.247)	(.399)
Promotion	-2.401***	-.725***	-.066
	(.369)	(.200)	(.064)
Animal	.940**	.425	-.008
	(.479)	(.259)	(.082)
Song	-.647*	-.344*	.128**
	(.357)	(.194)	(.061)
Mood: Emotional	.465	.126	-.064
	(.791)	(.430)	(.136)
Mood: Informational	-1.056*	-1.072***	-.451***
	(.615)	(.333)	(.106)
Mood: Funny	.520	.404**	.130**
	(.342)	(.186)	(.059)
Mood: Sexy	-1.848	-1.114	-.765**
	(1.850)	(1.009)	(.320)
Brand FE	Yes	Yes	Yes
N	18,861	18,991	18,973
R^2	.257	.313	.431
Adj. R^2	.152	.216	.351

Notes: The table reports the estimation results of Equation 3 in the main text using FGLS. Standard errors are reported in parentheses; ** p<.10, ** p<.05, *** p<.01.*

Table W18: Second-stage estimation results with arcsine transformation

	DV: estimated ad-creative fixed effects ($\hat{\delta}_i$)		
	(1)	(2)	(3)
	75%	50%	25%
Energy	.037***	.002	-.001
	(.012)	(.008)	(.003)
Duration: 30 sec	-.028***	-.093***	-.059***
	(.004)	(.002)	(.001)
Duration: 60 sec	-.052***	-.173***	-.169***
	(.010)	(.007)	(.003)
Duration: 90 sec	-.142***	-.277***	-.275***
	(.035)	(.023)	(.009)
Duration: > 90 sec	-.267***	-.416***	-.344***
	(.029)	(.019)	(.008)
Promotion	-.033***	-.018***	-.002*
	(.005)	(.003)	(.001)
Animal	.013**	.008*	-.001
	(.006)	(.004)	(.002)
Song	-.010**	-.009***	.001
	(.005)	(.003)	(.001)
Mood: Emotional	.005	.003	-.003
	(.010)	(.007)	(.003)
Mood: Informational	-.011	-.013**	-.008***
	(.008)	(.005)	(.002)
Mood: Funny	.010**	.010***	.004***
	(.004)	(.003)	(.001)
Mood: Sexy	-.012	-.004	-.012**
	(.024)	(.016)	(.006)
Brand FE	Yes	Yes	Yes
N	18,861	18,991	18,973
R^2	.267	.366	.587
Adj. R^2	.163	.277	.528

Notes: The table reports the estimation results of Equation 3 in the main text using FGLS with the arcsine transformation of ATR in the first stage. Standard errors are reported in parentheses; * *p<.10,* ** *p<.05,* *** *p<.01.*

Table W19: Second-stage estimation results with logit transformation

	DV: estimated ad-creative fixed effects ($\hat{\delta}_i$)		
	(1)	(2)	(3)
	75%	50%	25%
Energy	.143**	-.111	-.058
	(.070)	(.071)	(.111)
Duration: 30 sec†	-.194***	-1.067***	
	(.021)	(.022)	
Duration: 60 sec	-.327***	-1.686***	-1.924***
	(.059)	(.061)	(.068)
Duration: 90 sec	-.851***	-2.363***	-2.956***
	(.204)	(.210)	(.235)
Duration: > 90 sec	-1.455***	-3.185***	-3.529***
	(.171)	(.176)	(.201)
Promotion	-.205***	-.219***	-.104**
	(.028)	(.029)	(.043)
Animal	.086**	.082**	.033
	(.036)	(.037)	(.055)
Song	-.068**	-.104***	-.012
	(.027)	(.028)	(.043)
Mood: Emotional	.020	.014	-.046
	(.060)	(.062)	(.084)
Mood: Informational	-.044	-.074	-.115
	(.046)	(.049)	(.079)
Mood: Funny	.085***	.158***	.275***
	(.026)	(.027)	(.043)
Mood: Sexy	.018	.074	-.274
	(.140)	(.149)	(.309)
Brand FE	Yes	Yes	Yes
N	18,522	17,834	8,425
R^2	.276	.422	.466
Adj. R^2	.173	.339	.339

Notes: The table reports the estimation results of Equation 3 in the main text using FGLS with the logit transformation of ATR in the first stage. Standard errors are reported in parentheses; * *p<.10,* ** *p<.05,* *** *p<.01*

† For the 25% cutoff, "Duration: 30 sec" is the baseline because no observations used for the analysis are 15-second or shorter.

Table W20: Between-estimator second-stage estimation results (varying cutoff values)

	DV: estimated ad-creative fixed effects ($\hat{\delta}_i$)			
	(1)	(2)	(3)	(4)
Energy	-1.822***	.743	.807	2.197***
	(.638)	(.606)	(.613)	(.708)
Duration: 30 sec		7.963***	7.892***	8.410***
		(.174)	(.174)	(.218)
Duration: 60 sec		10.409***	10.648***	11.717***
		(.419)	(.430)	(.619)
Duration: 90 sec		7.723**	8.641**	4.733**
		(1.252)	(1.264)	(2.118)
Duration: > 90 sec		.176	.737	-.318
		(1.428)	(1.442)	(1.849)
Promotion		-2.342***	-2.291***	-.633**
		(.198)	(.198)	(.289)
Animal		1.277***	1.161***	.710*
		(.294)	(.295)	(.373)
Song		-.475**	-.549**	-.443
		(.226)	(.227)	(.279)
Mood: Emotional			.917*	.867
			(.509)	(.616)
Mood: Informational			-1.334***	-.377
			(.401)	(.483)
Mood: Funny			.470**	.433
			(.211)	(.267)
Mood: Sexy			-.004	-1.148
			(.976)	(1.449)
Constant	88.964***	85.386***	85.351***	
	(.221)	(.236)	(.244)	
Brand FE	No	No	No	Yes
N	19,015	19,005	18,943	18,943
R^2	<.001	.116	.117	.279
Adj. R^2	<.001	.116	.116	.176

Notes: The table reports the estimation results of Equation 3 in the main text using FGLS; energy is measured using the first 75% of ads' duration for 15-second or shorter ads, 50% for (15,30]-second ads, and 25% for <30-second ads; standard errors are reported in parentheses; * *p<.10,* ** *p<.05,* *** *p<.01.*

E.3. Accounting for Ad-wearout Effects

To account for potential ad-wearout effects, we first split out data into two time periods: period 1 (September 2015 - August 2016) and period 2 (September 2016 - August 2018). Then, we drop the ad-creatives in period 2 that appear at least once in period 1. By doing so, we attempt to address the initial condition problem since we do not observe the date on which an ad was first aired, which could have been before the start of our data. We drop about 24% of ad-creatives in period 2 and use the remaining observations for estimation. We first compute the cumulative number of prior viewership for each airing of an ad (hereafter, *cumulative viewership*). For instance, if an ad was first aired in period 2, the cumulative viewership for the ad for that particular insertion is zero. If the same ad was aired for the second time and a million viewership is recorded for the first airing, the cumulative viewership is a million.

We include this variable in the first-stage estimation of our between-estimator analysis (Equation 2 in the main text), together with all the fixed effects that we previously included. To be specific, we included log(1+cumulative viewership) to handle a potential nonlinear relationship between the variable and the ATR. The results are invariant when we include a quadratic term instead but the estimates are quite small since cumulative viewership can be large. We find that the coefficient for cumulative viewership is –.249 with the standard error of .005 ($p < .001$), which suggests that the ad-tuning rate systematically decreases as an ad is shown repetitively. This finding is consistent with general expectations about ad repetition effects.

Next, we compare the ad-creative fixed effects estimated with or without the ad-wearout effect. The correlation between the two sets of ad-creative fixed effects each of which is obtained from a model specification with or without the cumulative viewership variable is .998, which suggests that controlling for ad-wearout effect via cumulative viewership does not affect the *relative* magnitudes of the fixed effects across ad-creatives.

Nonetheless, Figure W16 shows that the ad-creative fixed effects estimated with the cumulative viewership variable tend to be greater than those estimated without the additional control variable (all points are above the blue 45-degree line). This makes intuitive sense when we consider the estimated ad-wearout effect is negative. The ad-creative fixed effects estimated without ad-wearout effects represent the quantities where the cumulative viewership is zero. Failing to account for this results in a systematic underestimation of ad-creative fixed effects

because we average out the declining quality of ad-creatives in terms of ad-tuning rate. This is what we basically see in the figure.

In Table W21, we compare the second-stage estimation results with or without cumulative viewership in the first stage. We find the two sets of estimates are largely quite similar and the estimate for the energy variable is still positive and statistically significant when the cumulative viewership variable was added to the first stage.

Figure W16: A scatterplot of estimated ad-creative fixed effects

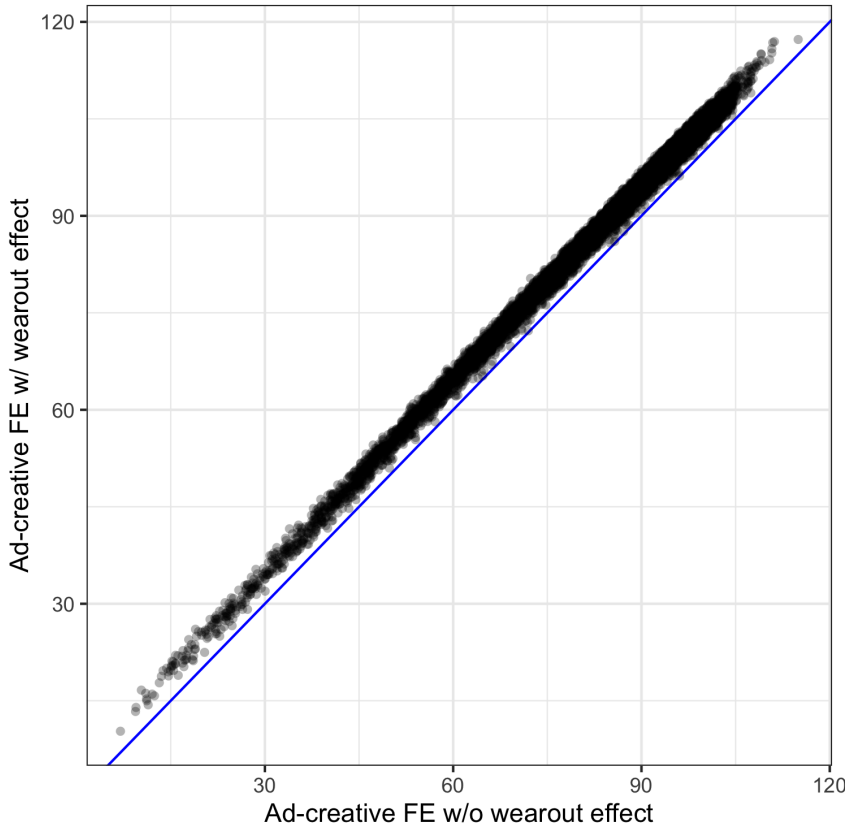


Table W21: Second-stage estimation results with controlling for ad-wearout effect

	DV: estimated ad-creative fixed effects ($\hat{\delta}_i$)	
	(1) w/o wearout	(2) w/ wearout
Energy	2.962***	2.290**
	(1.073)	(1.067)
Duration: 30 sec	-1.319***	-1.243***
	(.326)	(.324)
Duration: 60 sec	-2.990***	-3.068***
	(.956)	(.950)
Duration: 90 sec	-6.189*	-6.166*
	(3.269)	(3.250)
Duration: > 90 sec	-20.711***	-21.245***
	(2.696)	(2.686)
Promotion	-2.888***	-3.134***
	(.428)	(.425)
Animal	.732	.647
	(.535)	(.532)
Song	-.995**	-.985**
	(.423)	(.420)
Mood: Emotional	.260	.163
	(.941)	(.935)
Mood: Informational	-.916	-.892
	(.723)	(.719)
Mood: Funny	.005	.092
	(.403)	(.401)
Mood: Sexy	-2.405	-2.301
	(2.320)	(2.307)
Brand FE	Yes	Yes
N	13,959	13,959
R^2	.301	.303
Adj. R^2	.182	.185

Notes: The table reports the estimation results of Equation 3 in the main text using FGLS without or with the term for wearout effects is included in the first stage. Standard errors are reported in parentheses; *$p<.10$, **$p<.05$, ***$p<.01$.

WA-F. ADDITIONAL TABLES AND FIGURES

Table W22: Second-stage estimation results by product categories in all programs

	DV: estimated ad-creative fixed effects ($\hat{\delta}_{i\|g}$)			
	(1)	(2)	(3)	(4)
Energy				
x Apparel, Footwear and Accessories	1.869	1.133	.767	-5.821
	(2.172)	(2.146)	(2.145)	(5.275)
x Business and Legal	9.819***	8.724***	9.312***	8.719**
	(1.695)	(1.681)	(1.692)	(4.153)
x Education	-.225	-1.293	.020	3.073
	(4.005)	(3.974)	(3.987)	(12.366)
x Electronics and Communication	-8.260***	-5.924***	-6.300***	-.351
	(1.127)	(1.131)	(1.142)	(2.280)
x Food and Beverage	11.545***	7.095***	6.902***	2.908
	(1.070)	(1.080)	(1.089)	(2.200)
x Health and Beauty	6.478***	4.276***	4.857***	.419
	(1.244)	(1.237)	(1.252)	(2.741)
x Home and Real Estate	7.813***	5.392***	5.217***	1.242
	(1.520)	(1.507)	(1.512)	(3.677)
x Insurance	-3.697**	-3.994**	-4.676***	8.608*
	(1.798)	(1.781)	(1.799)	(4.467)
x Life and Entertainment	5.194***	4.654***	4.749***	9.668**
	(1.345)	(1.336)	(1.341)	(3.963)
x Pharmaceutical and Medical	.637	1.358	1.877	5.750
	(1.831)	(1.845)	(1.845)	(5.227)
x Politics, Government and Organization	6.367*	4.741	4.845	11.072
	(3.279)	(3.249)	(3.271)	(9.275)
x Restaurants	.725	-.590	-.808	-.274
	(1.429)	(1.421)	(1.426)	(3.625)
x Retail Stores	2.525**	5.696***	6.161***	4.206
	(1.166)	(1.219)	(1.222)	(2.718)
x Travel	10.559***	10.113***	10.013***	25.371***
	(2.742)	(2.718)	(2.716)	(8.165)
x Vehicles	-13.638***	-9.891***	-9.731***	7.050**
	(1.241)	(1.265)	(1.271)	(3.219)

Table continues in the next page

Table W22 (cont'd): Second-stage estimation results by product categories in all programs

	DV: estimated ad-creative fixed effects ($\hat{\delta}_{i\|g}$)			
	(1)	(2)	(3)	(4)
Duration: 30 sec		-1.770***	-1.853***	-1.810***
		(.242)	(.242)	(.279)
Duration: 60 sec		-5.215***	-4.713***	-3.724***
		(.559)	(.571)	(.789)
Duration: 90 sec		-2.790*	-1.104	-10.618***
		(1.668)	(1.682)	(2.711)
Duration: > 90 sec		-21.647***	-20.482***	-21.211***
		(1.772)	(1.787)	(2.320)
Promotion		-4.254***	-4.183***	-2.389***
		(.287)	(.287)	(.369)
Animal		1.606***	1.390***	.896*
		(.387)	(.388)	(.479)
Song		-.723**	-.837***	-.635*
		(.300)	(.302)	(.358)
Mood: Emotional			1.433**	.505
			(.671)	(.793)
Mood: Informational			-2.632***	-1.056*
			(.519)	(.615)
Mood: Funny			.927***	.484
			(.283)	(.342)
Mood: Sexy			-1.051	-1.927
			(1.279)	(1.850)
Constant	82.858***	85.170***	85.108***	
	(.283)	(.320)	(.330)	
Brand FE	No	No	No	Yes
N	18,933	18,923	18,861	18,861
R^2	.032	.058	.060	.259
Adj. R^2	.031	.056	.059	.152

Notes: The table reports the estimation results of Equation 3 in the main text using FGLS by product categories. Standard errors are reported in parentheses; * *p<.10,* ** *p<.05,* *** *p<.01.*

Figure W17: First-stage estimation results by program genres

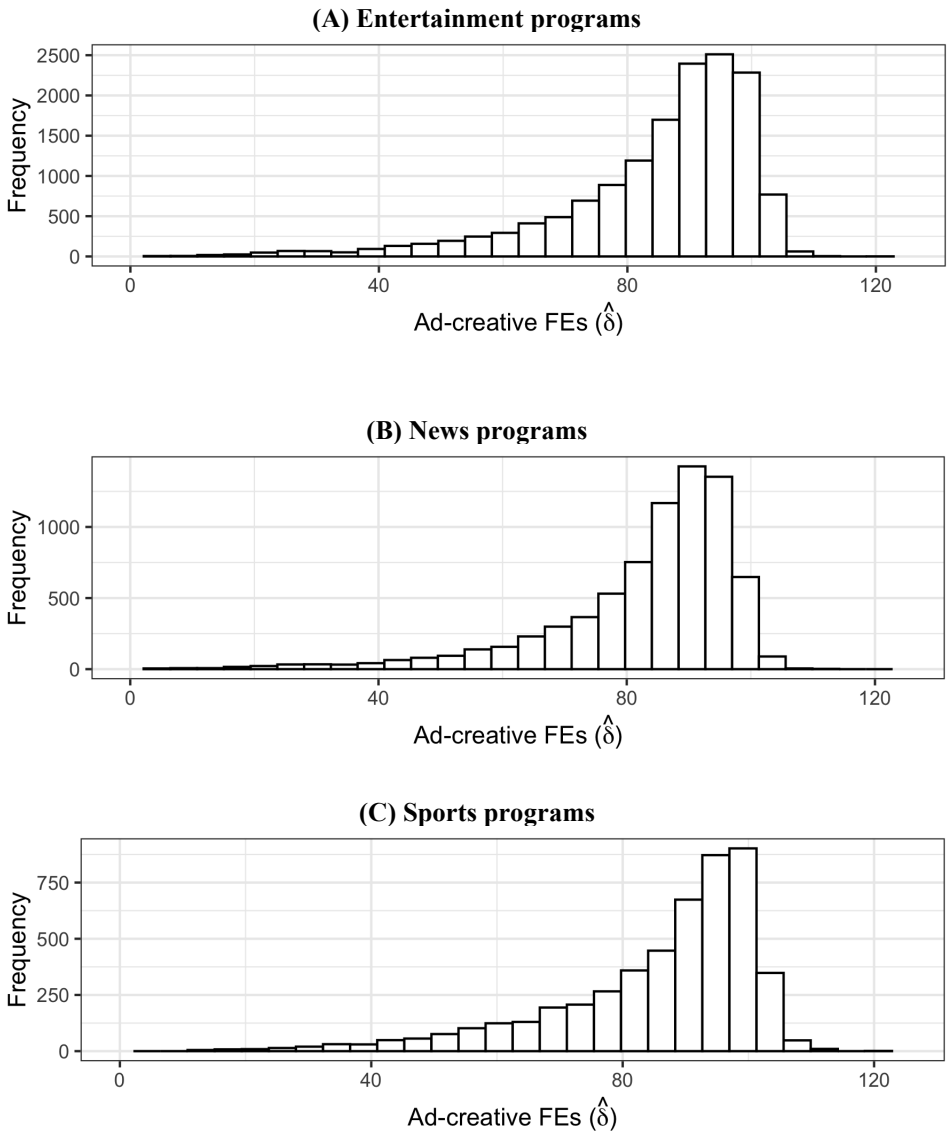


Table W23: Second-stage estimation results by product categories in entertainment programs

	DV: estimated ad-creative fixed effects ($\hat{\delta}_{i|g}$)			
	(1)	(2)	(3)	(4)
Energy				
x Apparel, Footwear and Accessories	-2.660	-4.072	-4.261	-8.921
	(3.041)	(3.008)	(3.002)	(7.257)
x Business and Legal	4.542**	3.601*	4.661**	1.070
	(2.165)	(2.148)	(2.177)	(5.350)
x Education	3.392	2.689	3.529	-14.060
	(6.388)	(6.348)	(6.359)	(17.152)
x Electronics and Communication	-10.551***	-7.760***	-7.939***	-1.004
	(1.272)	(1.290)	(1.302)	(2.593)
x Food and Beverage	12.839***	8.339***	8.326***	5.549**
	(1.210)	(1.222)	(1.232)	(2.476)
x Health and Beauty	8.716***	6.766***	7.422***	1.106
	(1.347)	(1.339)	(1.354)	(2.864)
x Home and Real Estate	9.600***	7.405***	7.215***	1.480
	(1.697)	(1.684)	(1.689)	(4.007)
x Insurance	-5.054**	-3.816*	-4.168**	8.050
	(2.040)	(2.023)	(2.042)	(5.062)
x Life and Entertainment	6.163***	5.628***	5.878***	9.580*
	(1.691)	(1.684)	(1.687)	(5.160)
x Pharmaceutical and Medical	2.747	3.355*	3.750*	9.558*
	(1.924)	(1.944)	(1.943)	(5.393)
x Politics, Government and Organization	4.290	4.184	4.126	10.843
	(4.391)	(4.349)	(4.397)	(12.464)
x Restaurants	3.023*	1.427	1.414	-1.489
	(1.589)	(1.583)	(1.589)	(4.031)
x Retail Stores	3.793***	5.904***	6.484***	2.508
	(1.267)	(1.335)	(1.339)	(2.954)
x Travel	13.381***	12.317***	12.372***	34.235***
	(3.314)	(3.295)	(3.289)	(9.879)
x Vehicles	-18.720***	-14.384***	-13.977***	11.210***
	(1.518)	(1.564)	(1.574)	(4.006)

Table continues in the next page

Table W23 (cont'd): Second-stage estimation results by product categories in entertainment programs

	DV: estimated ad-creative fixed effects ($\hat{\delta}_{i\|g}$)			
	(1)	(2)	(3)	(4)
Duration: 30 sec		-2.177***	-2.271***	-2.106***
		(.274)	(.274)	(.311)
Duration: 60 sec		-5.463***	-4.966***	-4.417***
		(.634)	(.651)	(.995)
Duration: 90 sec		-1.436	.234	-13.665***
		(1.822)	(1.840)	(3.352)
Duration: > 90 sec		-24.026***	-22.983***	-24.941***
		(1.952)	(1.967)	(2.639)
Promotion		-3.375***	-3.337***	-1.696***
		(.324)	(.324)	(.416)
Animal		1.988***	1.787***	1.378**
		(.426)	(.427)	(.539)
Song		-.465	-.594*	-.672*
		(.335)	(.338)	(.403)
Mood: Emotional			1.930**	1.450
			(.795)	(.954)
Mood: Informational			-2.096***	-.659
			(.565)	(.678)
Mood: Funny			.888***	.296
			(.319)	(.387)
Mood: Sexy			-.841	-2.805
			(1.379)	(2.002)
Constant	84.346***	86.513***	86.399***	
	(.319)	(.355)	(.369)	
Brand FE	No	No	No	Yes
N	14,701	14,692	14,637	14,637
R^2	.045	.070	.073	.268
Adj. R^2	.044	.069	.071	.163

Notes: The table reports the estimation results of Equation 3 in the main text using FGLS by product categories. Standard errors are reported in parentheses; * *p<.10,* ** *p<.05,* *** *p<.01.*

Table W24: Second-stage estimation results by product categories in news programs

	DV: estimated ad-creative fixed effects ($\hat{\delta}_{i|g}$)			
	(1)	(2)	(3)	(4)
Energy				
x Apparel, Footwear and Accessories	-3.779	-3.197	-3.136	-12.384
	(4.216)	(4.202)	(4.199)	(10.051)
x Business and Legal	10.435***	10.384***	10.451***	7.114
	(2.735)	(2.728)	(2.730)	(7.167)
x Education	-2.312	-3.054	-2.831	.229
	(5.498)	(5.476)	(5.467)	(17.303)
x Electronics and Communication	-8.268***	-5.070**	-5.183**	7.775
	(2.145)	(2.184)	(2.198)	(5.065)
x Food and Beverage	10.448***	7.025***	7.082***	2.995
	(1.673)	(1.714)	(1.726)	(3.353)
x Health and Beauty	3.484**	2.047	2.696	-1.842
	(1.753)	(1.768)	(1.791)	(3.611)
x Home and Real Estate	3.517*	1.867	1.727	-4.657
	(2.121)	(2.122)	(2.132)	(4.936)
x Insurance	-3.894*	-4.204*	-4.642*	9.791
	(2.334)	(2.342)	(2.406)	(6.003)
x Life and Entertainment	3.445	5.679	5.676	10.966
	(3.740)	(3.751)	(3.747)	(11.181)
x Pharmaceutical and Medical	.697	.211	.165	10.976*
	(2.153)	(2.202)	(2.202)	(5.616)
x Politics, Government and Organization	-1.562	-2.570	-2.812	11.483
	(4.295)	(4.278)	(4.284)	(11.993)
x Restaurants	2.294	2.868	2.887	2.516
	(2.719)	(2.732)	(2.737)	(7.321)
x Retail Stores	-.225	3.046*	3.189*	2.976
	(1.648)	(1.804)	(1.806)	(3.890)
x Travel	9.823**	10.213**	10.543***	21.059
	(4.068)	(4.051)	(4.052)	(14.016)
x Vehicles	-17.652***	-14.292***	-14.541***	-10.152
	(2.446)	(2.492)	(2.503)	(7.266)

Table continues in the next page

Table W24 (cont'd): Second-stage estimation results by product categories in news programs

	DV: estimated ad-creative fixed effects ($\hat{\delta}_{i\|g}$)			
	(1)	(2)	(3)	(4)
Duration: 30 sec		-1.474***	-1.528***	-1.471***
		(.375)	(.375)	(.439)
Duration: 60 sec		-1.846**	-1.645*	-2.811*
		(.854)	(.879)	(1.578)
Duration: 90 sec		-1.373	-.124	-10.905***
		(1.968)	(1.995)	(3.978)
Duration: > 90 sec		-13.393***	-13.173***	-11.797***
		(2.764)	(2.779)	(4.253)
Promotion		-3.365***	-3.324***	-1.900***
		(.469)	(.470)	(.651)
Animal		2.353***	2.240***	1.736**
		(.542)	(.543)	(.741)
Song		-.533	-.551	-.855
		(.497)	(.500)	(.615)
Mood: Emotional			1.088	1.133
			(.931)	(1.168)
Mood: Informational			-.685	-.776
			(.737)	(.861)
Mood: Funny			.491	-.485
			(.463)	(.577)
Mood: Sexy			-2.764	-2.659
			(2.021)	(2.768)
Constant	82.666***	84.001***	83.941***	
	(.419)	(.476)	(.490)	
Brand FE	No	No	No	Yes
N	7,558	7,558	7,537	7,537
R^2	.025	.040	.041	.295
Adj. R^2	.023	.037	.038	.148

Notes: The table reports the estimation results of Equation 3 in the main text using FGLS by product categories. Standard errors are reported in parentheses; * *p<.10,* ** *p<.05,* *** *p<.01.*

Table W25: Second-stage estimation results by product categories in sports programs

	DV: estimated ad-creative fixed effects ($\hat{\delta}_{i|g}$)			
	(1)	(2)	(3)	(4)
Energy				
x Apparel, Footwear and Accessories	-1.730	-2.435	-2.531	-6.754
	(3.402)	(3.382)	(3.390)	(8.591)
x Business and Legal	14.866***	12.708***	12.342***	9.659
	(2.969)	(2.952)	(2.964)	(6.905)
x Education	-.105	-3.319	-2.862	8.693
	(10.070)	(9.991)	(10.003)	(37.682)
x Electronics and Communication	-7.991***	-5.144**	-5.507***	-4.210
	(2.100)	(2.102)	(2.123)	(3.800)
x Food and Beverage	4.594**	1.558	1.284	-4.740
	(2.112)	(2.157)	(2.171)	(4.928)
x Health and Beauty	2.077	-.742	-1.164	5.824
	(3.434)	(3.444)	(3.468)	(10.417)
x Home and Real Estate	3.889	.494	.170	-3.478
	(3.626)	(3.606)	(3.619)	(11.534)
x Insurance	-2.842	-5.473**	-6.307**	1.587
	(2.706)	(2.699)	(2.765)	(6.626)
x Life and Entertainment	.678	.621	.724	-4.065
	(2.549)	(2.532)	(2.551)	(7.953)
x Pharmaceutical and Medical	-20.582***	-23.137***	-23.151***	-16.058
	(5.412)	(5.511)	(5.550)	(23.303)
x Politics, Government and Organization	17.342**	14.183**	14.758**	15.715
	(6.962)	(6.904)	(6.923)	(22.582)
x Restaurants	-7.450***	-6.987***	-7.297***	-.014
	(2.511)	(2.532)	(2.546)	(6.241)
x Retail Stores	-5.106	-2.166	-2.314	-18.993*
	(3.576)	(3.577)	(3.588)	(10.702)
x Travel	3.842	5.061	4.873	-18.145
	(5.218)	(5.168)	(5.175)	(17.568)
x Vehicles	-10.039***	-6.678***	-6.730***	.346
	(2.102)	(2.115)	(2.133)	(5.212)

Table continues in the next page

Table W25 (cont'd): Second-stage estimation results by product categories in sports programs

	DV: estimated ad-creative fixed effects ($\hat{\delta}_{i\|g}$)			
	(1)	(2)	(3)	(4)
Duration: 30 sec		-1.015*	-.990*	-1.337*
		(.552)	(.553)	(.710)
Duration: 60 sec		-2.034	-1.797	-2.021
		(1.354)	(1.374)	(1.699)
Duration: 90 sec		2.832	2.843	-3.618
		(4.439)	(4.448)	(6.759)
Duration: > 90 sec		3.987	3.549	-13.880
		(13.633)	(13.720)	(25.090)
Promotion		-6.377***	-6.461***	-3.532***
		(.615)	(.617)	(.766)
Animal		1.421	1.373	1.033
		(.869)	(.872)	(1.025)
Song		-1.415**	-1.349**	-.354
		(.581)	(.585)	(.717)
Mood: Emotional			-1.368	-4.339***
			(1.343)	(1.619)
Mood: Informational			.532	1.428
			(1.686)	(1.992)
Mood: Funny			.623	.404
			(.539)	(.699)
Mood: Sexy			8.904	9.066
			(9.910)	(14.429)
Constant	86.474***	88.599***	88.529***	
	(.596)	(.730)	(.754)	
Brand FE	No	No	No	Yes
N	4,939	4,938	4,934	4,934
R^2	.029	.052	.053	.281
Adj. R^2	.026	.048	.048	.104

Notes: The table reports the estimation results of Equation 3 in the main text using FGLS by product categories. Standard errors are reported in parentheses; * *p<.10,* ** *p<.05,* *** *p<.01.*

WEB APPENDIX REFERENCES

Breiman, Leo (2001), "Random Forests," *Machine Learning*, 45 (1), 5-32.

Chen, Tianqi, and Carlos Guestrin (2016), "XGBoost: A Scalable Tree Boosting System," In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery,* 785-794.

Darshna, Parmar (2018), "Music Recommendation based on Content and Collaborative Approach & Reducing Cold Start Problem," In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 1033-37, IEEE.

Friedman, Jerome H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29 (5):1189–1232.

Jiang, Dan-Ning, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai (2002), "Music Type Classification by Spectral Contrast Feature," In *ICME'02*. vol. 1, pp. 113-116. IEEE.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015), "Deep Learning," *Nature,* 521 (7553), 436-444.

Millecamp, Martijn, Nyi Htun, Yucheng Jin, and Katrien Verbert (2018), "Controlling Spotify Recommendations: Effects of Personal Characteristics on Music Recommender User Interfaces," In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization,* 101-109.

Nielsen, Didrik (2016), "Tree Boosting with XGBoost–Why Does XGBoost Win 'Every' Machine Learning Competition?" *Norwegian University of Science and Technology*.

Ripley, Brian D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press.

Schwenzow, Jasper, Jochen Hartmann, Amos Schikowsky, and Mark Heitmann (2021), "Understanding Videos at Scale: How to Extract Insights for Business Research," *Journal of Business Research*, 123 (February), 367-379.

Stevens, Stanley Smith, John Volkmann, and Edwin B. Newman (1937). "A Scale for the Measurement of the Psychological Magnitude Pitch," *The Journal of the Acoustical Society of America* 8 (3), 185-190.

Tibshirani, Robert (1997), "The Lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine*, 16 (4), 385-395.