Data-intensive Innovation and the State: Evidence from AI Firms in China

Martin Beraja
MIT

David Yang
Harvard

Noam Yuchtman
LSE

SED
July, 2021

Research assistants: Haoran Gao, Shiyun Hu, Andrew Kao, Shuhao Lu, Junxi Liu, Shengqi Ni, Wenwei Peng, Yucheng Quan, Linchuan Xu, Peilin Yang, and Guoli Yin

Motivation: government data as input in AI innovation

- AI innovation is **data-intensive**
  - Many recent AI advances made with decades-old algorithms applied to newly available big data

Motivation: government data as input in AI innovation

- AI innovation is **data-intensive**
  - Many recent AI advances made with decades-old algorithms applied to newly available big data
- Literature has focused on how data collected by **private** firms shapes AI innovation (Agrawal et al., 2019; Jones and Tonetti, 2020)

Motivation: government data as input in AI innovation

- AI innovation is **data-intensive**
  - Many recent AI advances made with decades-old algorithms applied to newly available big data
- Literature has focused on how data collected by **private** firms shapes AI innovation (Agrawal et al., 2019; Jones and Tonetti, 2020)
- Yet, throughout history, **states** have also collected massive quantities of data (Scott, 1998)
- The state has a large role in many areas
  - Public security, health care, education, basic science...

$\Longrightarrow$ **Government data** can exceed privately-collected data in magnitude/scope; or lack good substitutes altogether

Motivation: China's facial recognition AI sector

- A common way in which AI firms **gain access** to valuable government data is by **providing services** to the state

Motivation: China's facial recognition AI sector

- A common way in which AI firms **gain access** to valuable government data is by **providing services** to the state
- Think about **facial recognition AI firms in China**...
  - Train algorithms with, e.g., video streams of faces from many angles
  - The state's public security units collect this form of data through their surveillance apparatus, and contract AI firms for services
  - AI firms gaining access to surveillance data can use it to train algorithms and develop software

This paper

Does access to **government data** when providing AI services to the state stimulate **commercial** AI innovation?

This paper

Does access to **government data** when providing AI services to the state stimulate **commercial** AI innovation?

The mechanism(s)

1. If gov't data and algorithms are **sharable** across uses, they can be used to develop AI products for commercial markets
   (e.g., a facial recognition platform for retail stores)

2. Firms may **learn** to manage and utilize large datasets too

$\Longrightarrow$ a procurement contract with access to gov't data can fuel commercial innovation, overcoming **crowd-out** from the contract

This paper

Does access to **government data** when providing AI services to the state stimulate **commercial** AI innovation?

The mechanism(s)

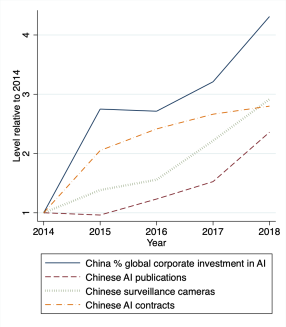
1. If gov't data and algorithms are **sharable** across uses, they can be used to develop AI products for commercial markets (e.g., a facial recognition platform for retail stores)
2. Firms may **learn** to manage and utilize large datasets too

$\Longrightarrow$ a procurement contract with access to gov't data can fuel commercial innovation, overcoming **crowd-out** from the contract

Evidence of this in China's facial recognition AI sector

Two implications

1. Access to gov't data contributed to Chinese firms' emergence as leading innovators in facial recognition AI
   - Indeed, this has coincided with the expansion of the government's procurement of AI and surveillance capacity



Two implications

1. Access to gov't data contributed to Chinese firms' emergence as leading innovators in facial recognition AI
   - Indeed, this has coincided with the expansion of the government's procurement of AI and surveillance capacity
2. Novel role for the state in data-intensive economies
   - So far, emphasis on the regulation of privately-collected data due to antitrust or privacy concerns (Tirole, 2020; Aridor et al., 2020)
   - AI procurement and policies of gov't data collection and provision could, **whether intentionally or not**, stimulate and shape the direction of innovation in a range of sectors

Empirical challenges

Would like to compare software output changes after receipt of gov't procurement contracts giving access to more v. less data

Empirical challenges

Would like to compare software output changes after receipt of gov't procurement contracts giving access to more v. less data

Data challenges

1. Dataset linking AI firms to govt. contracts did not exist
2. Dataset on AI firms' software did not exist (our measure of *product innovation*). Also, critical for us to classify by use (commercial or not)
3. No available direct measures of firm-level use of gov't data

Empirical challenges

Would like to compare software output changes after receipt of gov't procurement contracts giving access to more v. less data

Data challenges

1. Dataset linking AI firms to govt. contracts did not exist
2. Dataset on AI firms' software did not exist (our measure of *product innovation*). Also, critical for us to classify by use (commercial or not)
3. No available direct measures of firm-level use of gov't data

Identification challenges

1. Non-random assignment of gov't contracts
2. Contracts work through other mechanisms unrelated to data

Data 1: linking AI firms to govt. contracts

1. Identify all facial recognition AI **firms**

    - 7,837 firms
    - Two sources: Tianyancha (People's Bank of China) and PitchBook (Morningstar)
    - Include: *(i)* firms specialized in facial recognition AI (e.g., Yitu); *(ii)* hardware firms that devote substantial resources to develop AI software (e.g., Hik-Vision); *(iii)* facial recognition AI units of large tech conglomerates (e.g., Baidu AI)

Data 1: linking AI firms to govt. contracts

1. Identify all facial recognition AI **firms**
   - 7,837 firms
   - Two sources: Tianyancha (People's Bank of China) and PitchBook (Morningstar)
   - Include: *(i)* firms specialized in facial recognition AI (e.g., Yitu); *(ii)* hardware firms that devote substantial resources to develop AI software (e.g., Hik-Vision); *(iii)* facial recognition AI units of large tech conglomerates (e.g., Baidu AI)

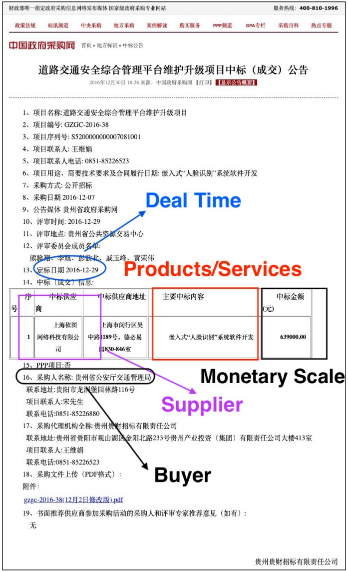
2. Obtain universe of **government contracts**
   - 2,997,105 contracts
   - Source: Chinese Govt. Procurement Database (Ministry of Finance)

Data 1: linking AI firms to govt. contracts

1. Identify all facial recognition AI **firms**
   - 7,837 firms
   - Two sources: Tianyancha (People's Bank of China) and PitchBook (Morningstar)
   - Include: *(i)* firms specialized in facial recognition AI (e.g., Yitu); *(ii)* hardware firms that devote substantial resources to develop AI software (e.g., Hik-Vision); *(iii)* facial recognition AI units of large tech conglomerates (e.g., Baidu AI)

2. Obtain universe of **government contracts**
   - 2,997,105 contracts
   - Source: Chinese Govt. Procurement Database (Ministry of Finance)

3. Link government **buyers** to AI **suppliers**


财政部唯一指定政府采购信息网络发布媒体 国家级政府采购专业网站 服务热线：400-810-1996

政策法规 标讯频道 中央采购 地方采购 案例解读 购买服务 PPP频道 GPA专栏 采购百科 热点专题

中国政府采购网 首页 » 地方标讯 » 中标公告

道路交通安全综合管理平台维护升级项目中标（成交）公告

2016年12月30日 16:26 来源：中国政府采购网 【打印】【显示公告概要】

1、项目名称:道路交通安全综合管理平台维护升级项目
2、项目编号: GZGC-2016-38
3、项目序列号: S5200000000007081001
4、项目联系人: 王维娟
5、项目联系人电话: 0851-85226523
6、项目用途、简要技术要求及合同履行日期: 嵌入式"人脸识别"系统软件开发
7、采购方式: 公开招标
8、采购日期 2016-12-07
9、公告媒体 贵州省政府采购网
10、评审时间: 2016-12-29
11、评审地点: 贵州省公共资源交易中心
12、评审委员会成员名单:
熊验翔、李旭、彭歆北、戚玉峰、黄荣伟
13、定标日期 2016-12-29
14、中标（成交）信息:

序号	中标供应商	中标供应商地址	主要中标内容	中标金额(元)
1	上海依图网络科技有限公司	上海市闵行区吴中路1189号，德必易园830-846室	嵌入式"人脸识别"系统软件开发	639000.00

15、PPP项目:否
16、采购人名称: 贵州省公安厅交通管理局
联系地址:贵阳市龙洞堡园林路116号
项目联系人:宋先生
联系电话:0851-85226880
17、采购代理机构全称:贵州贵财招标有限责任公司
联系地址:贵州省贵阳市观山湖区金阳北路233号贵州产业投资（集团）有限责任公司大楼413室
项目联系人:王维娟
联系电话:0851-85226523
18、采购文件上传（PDF格式）:
附件:
gzgc-2016-38(12月2日修改版).pdf
19、书面推荐供应商参加采购活动的采购人和评审专家推荐意见（如有）:
无

贵州贵财招标有限责任公司

Data 2: AI firms' software production

Registered with Min. of Industry and Information Technology

- Validation exercise: check against IPO Prospectus of MegVii

Data 2: AI firms' software production

Registered with Min. of Industry and Information Technology

- Validation exercise: check against IPO Prospectus of MegVii

Categorize by intended customers:

1. **Commercial:** e.g., *visual recognition system for smart retail*;
2. **Government:** e.g., *smart city — real time monitoring system on main traffic routes*;
3. General: e.g., *a synchronization method for multi-view cameras based on FPGA chips*.

Categorization: analyze text using machine learning

- Recurrent Neural Network (RNN) model using tensorflow
  - Corpus: 13,000 manually labeled software programs
  - Word-embedding: converted sentences to vectors based on word frequencies and used the words from full datasets as dictionary
  - Long Short-Term Memory (LSTM) algorithm: 2 layers of 32 nodes
  - 90% of corpus for training, 10% for validating
  - 10,000 training cycles are run for gradient descent on loss function
- Results robust to perturbing parameters of learning model

Data 3: measuring access to government data

Within AI public security contracts: variation in the data collection capacity of the public security agency's local surveillance network

1. Identify non-AI contracts: police department purchases of street cameras
2. Measure quantity of advanced cameras in a prefecture at a given time
3. Categorize public security contracts as coming from "high" or "low" camera capacity prefectures

Baseline empirical strategy

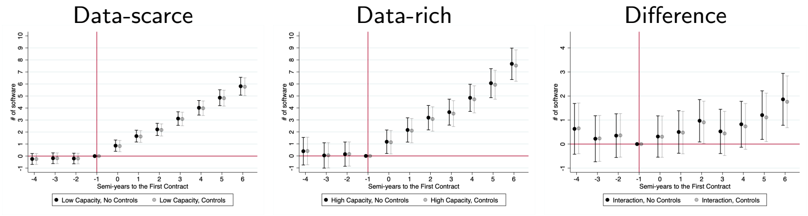
- **Triple diffs:** compare cumulative software releases before and after firms received 1st data-rich contracts, relative to the data-scarce ones

$$y_{it} = \sum_T \beta_{1T} T_{it} Data_i + \sum_T \beta_{2T} T_{it} + \alpha_t + \gamma_i + \sum_T \beta_{3T} T_{it} X_i + \epsilon_{it}$$

  - T_{it}: 1 if, at time t, T semi-years have passed before/since firm i received 1st contract
  - $Data_i$: 1 if firm i receives "data rich" contract (i.e., from "high" camera capacity prefecture at time of contract receipt)
  - X_i controls for pre-contract firm characteristics: age, size (cap), and software production

Public security contract “richer in data” & firm innovation

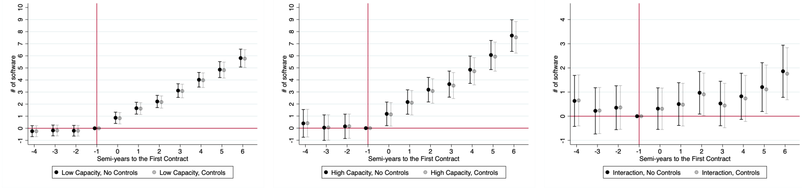
Commercial use cumulative software releases



Public security contract "richer in data" & firm innovation

Commercial use cumulative software releases

Data-scarce | Data-rich | Difference

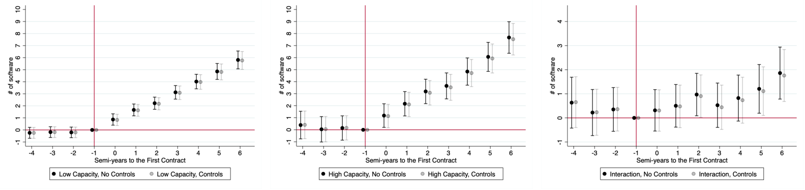


Magnitude: 2 new software products over 3 years

Public security contract “richer in data” & firm innovation

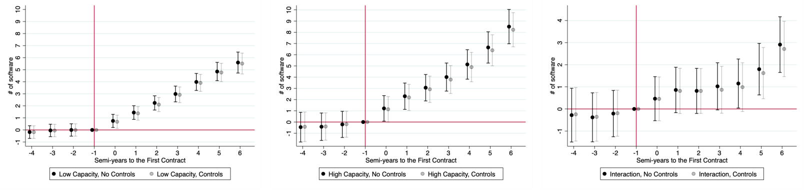
Commercial use cumulative software releases

Data-scarce | Data-rich | Difference



Government use cumulative software releases

Data-scarce | Data-rich | Difference



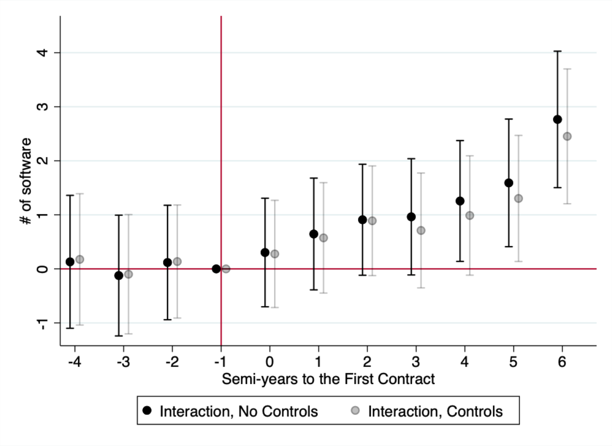
Commercial innovation overcomes crowd-out of inputs by gov’t

Evaluating alternative hypotheses

1. **Selection** at a given time differs by contract
   - Firm controls. No differential pre-contract levels/trends of software
2. **Terms and tasks** differ by contract ▸ Language distance
   - Descriptions of data-rich and -scarce contracts are similar in content
   - Similar govt soft produced after data-rich and -scarce contracts too
3. **Importance of capital** differs by contract ▸ Capital
   - Control for time-period×: pre-contract market cap or amount of external financing, and monetary value of contract
4. **Signals** differ by contract ▸ Signals
   - Subsamples of firms: (i) from a *mother* firm that has already received contract, or (ii) receiving a 2nd data-rich contract
5. **Govt connections** or **opportunities** differ by contract ▸ Local
   - Drop contracts with Beijing/Shangai or firm's home province.
   - Control for time-period×GDP-per-cap

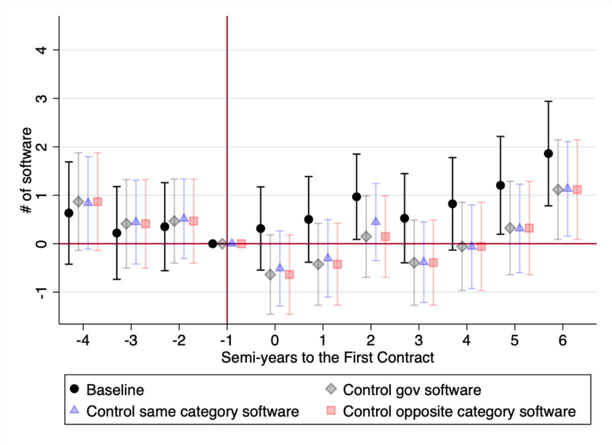
Additional evidence for our mechanism(s)

Data-complementary software (e.g., storage/transmission) differentially increases after data-rich contract. **Learning?**



Additional evidence for our mechanism(s)

- Include **pre-contract AI production** interacted with **Time period fixed-effects**. (Over)controls for **learning potential**
- Baseline estimate still positive, but halves in magnitude. **Direct effect due to sharability of data/algorithms?**



Contributions to literature

1. To the literature on the economics of AI and data (e.g., Aghion et al., 2017; Agrawal et al., 2018; Farboodi et al., 2019; Jones and Tonetti, 2019)

    - Highlight the role of **government data** in shaping commercial AI innovation, and the **sharability of data/algorithms** within the firm

Contributions to literature

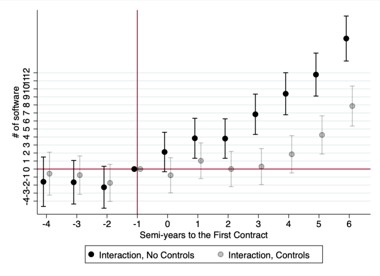
1. To the literature on the economics of AI and data (e.g., Aghion et al., 2017; Agrawal et al., 2018; Farboodi et al., 2019; Jones and Tonetti, 2019)
   - Highlight the role of **government data** in shaping commercial AI innovation, and the **sharability of data/algorithms** within the firm
2. To the literature on industrial and innovation policies (e.g., Rodrik, 2007; Lane, 2020; Bloom et al., 2019)
   - Government data provision to firms can act as an innovation policy, **whether intentionally or not**
   - Mechanisms **similar** to other government policies (e.g., learning spillovers from space exploration) but **distinct** too (direct effect of sharability)

Contributions to literature

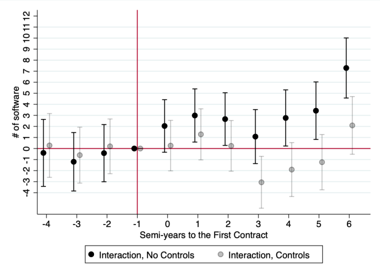
1. To the literature on the economics of AI and data (e.g., Aghion et al., 2017; Agrawal et al., 2018; Farboodi et al., 2019; Jones and Tonetti, 2019)
   - Highlight the role of **government data** in shaping commercial AI innovation, and the **sharability of data/algorithms** within the firm
2. To the literature on industrial and innovation policies (e.g., Rodrik, 2007; Lane, 2020; Bloom et al., 2019)
   - Government data provision to firms can act as an innovation policy, **whether intentionally or not**
   - Mechanisms **similar** to other government policies (e.g., learning spillovers from space exploration) but **distinct** too (direct effect of sharability)
3. To the literature on the rise of China emphasizing the role of the state (e.g., Lau et al., 2000; Brandt and Rawski, 2008; Song et al., 2011)
   - Highlight the role of the **surveillance apparatus** in commercial innovation
   - *Next project:* **AI-tocracy**. Alignment between innovation and autocracy? Contrasts with e.g., North (1991); Acemoglu and Robinson (2006, 2012)

Appendix

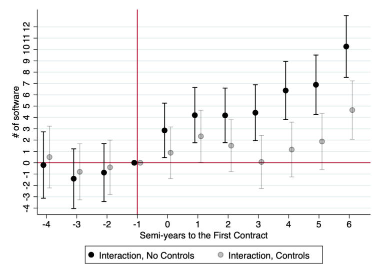
(a) Government (for video-AI)



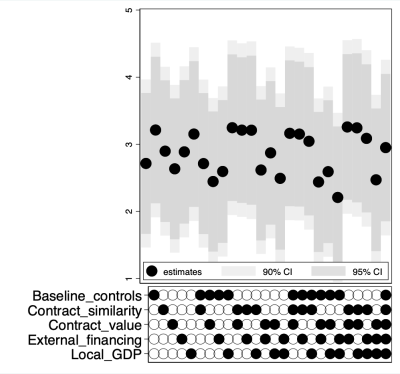
(b) Commercial (for video-AI)



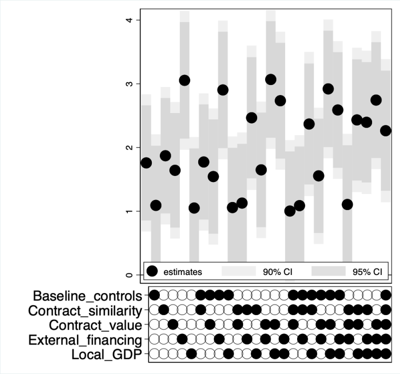
(c) Data-complementary (for video-AI)



(a) Government



(b) Commercial



Back

Table A.11: Scale effects and learning-by-doing

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A: Baseline			
4 Semiyears Before	-0.177	-0.239	-0.310
	(0.268)	(0.231)	(0.270)
6 Semiyears After	5.595***	5.811***	6.383***
	(0.444)	(0.378)	(0.443)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130
	(0.620)	(0.539)	(0.627)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***
	(0.642)	(0.550)	(0.644)
Panel B: Control for government pre-contract software production			
4 Semiyears Before	0.138	-0.076	-0.081
	(0.233)	(0.220)	(0.252)
6 Semiyears After	1.769***	3.846***	3.652***
	(0.386)	(0.362)	(0.415)
4 Semiyears Before × High Capacity	0.170	0.869*	0.489
	(0.538)	(0.514)	(0.586)
6 Semiyears After × High Capacity	1.477***	1.116**	1.722***
	(0.556)	(0.525)	(0.602)
Panel C: Control for same category pre-contract software production			
4 Semiyears Before	0.138	0.034	-0.047
	(0.233)	(0.209)	(0.253)
6 Semiyears After	1.769***	2.577***	3.173***
	(0.386)	(0.344)	(0.418)
4 Semiyears Before × High Capacity	0.170	0.841*	0.361
	(0.538)	(0.487)	(0.589)
6 Semiyears After × High Capacity	1.477***	1.132**	2.013***
	(0.556)	(0.498)	(0.605)
Panel D: Control for opposite category pre-contract software production			
4 Semiyears Before	0.080	-0.076	-0.061
	(0.250)	(0.220)	(0.256)
6 Semiyears After	2.399***	3.846***	3.474***
	(0.416)	(0.362)	(0.423)
4 Semiyears Before × High Capacity	-0.078	0.869*	0.302
	(0.579)	(0.514)	(0.596)
6 Semiyears After × High Capacity	2.231***	1.116**	2.111***
	(0.599)	(0.525)	(0.612)

Table A.12: Effects of 2nd public security contracts

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A: Baseline			
4 Semiyears Before	-0.177	-0.239	-0.310
	(0.268)	(0.231)	(0.270)
6 Semiyears After	5.595***	5.811***	6.383***
	(0.444)	(0.378)	(0.443)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130
	(0.620)	(0.539)	(0.627)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***
	(0.642)	(0.550)	(0.644)
Panel B: Sample — not first contract within mother firm			
4 Semiyears Before	-0.078	-0.431	-0.184
	(0.213)	(0.362)	(0.283)
6 Semiyears After	4.606***	6.730***	6.370***
	(0.332)	(0.557)	(0.438)
4 Semiyears Before × High Capacity	1.035	1.047	0.820
	(0.786)	(1.384)	(1.081)
6 Semiyears After × High Capacity	2.753***	1.975*	1.024
	(0.710)	(1.200)	(0.947)
Panel C: Sample — second contract within subsidiary firm			
4 Semiyears Before	-1.577*	2.214***	2.015***
	(0.916)	(0.656)	(0.697)
6 Semiyears After	8.533***	7.856***	13.538***
	(1.430)	(1.025)	(1.088)
4 Semiyears Before × High Capacity	1.090	-1.943**	-1.819*
	(1.287)	(0.923)	(0.980)
6 Semiyears After × High Capacity	29.042***	2.876**	17.833***
	(1.881)	(1.349)	(1.432)

Back

Table A.13: Robustness — firm geography

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A: Baseline			
4 Semiyears Before	-0.177	-0.239	-0.310
	(0.268)	(0.231)	(0.270)
6 Semiyears After	5.595***	5.811***	6.383***
	(0.444)	(0.378)	(0.443)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130
	(0.620)	(0.539)	(0.627)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***
	(0.642)	(0.550)	(0.644)
Panel B: Drop Beijing, Shanghai			
4 Semiyears Before	-0.179	-0.242	-0.277
	(0.264)	(0.166)	(0.249)
6 Semiyears After	5.511***	5.873***	6.286***
	(0.423)	(0.264)	(0.397)
4 Semiyears Before × High Capacity	-0.114	0.763*	0.235
	(0.634)	(0.404)	(0.603)
6 Semiyears After × High Capacity	2.983***	1.118***	2.863***
	(0.641)	(0.403)	(0.605)
Panel C: Firm based outside contract province			
4 Semiyears Before	-0.195	-0.165	-0.293
	(0.209)	(0.245)	(0.218)
6 Semiyears After	5.254***	5.862***	6.153***
	(0.333)	(0.387)	(0.346)
4 Semiyears Before × High Capacity	-0.053	0.721	0.177
	(0.555)	(0.658)	(0.586)
6 Semiyears After × High Capacity	2.365***	2.747***	2.815***
	(0.542)	(0.636)	(0.567)

Back