

Mass Digitization of Chinese Court Decisions

Rachel Stern and Luke Sanford

joint work with:

Benjamin Liebman & Alice Wang (Columbia Law School)
Molly Roberts (UC San Diego)

29 million Chinese court cases



29 million Chinese court cases



- Unprecedented transparency: Required to upload since 2014

29 million Chinese court cases



- ▶ **Unprecedented transparency:** Required to upload since 2014
- ▶ Why?

29 million Chinese court cases



- ▶ **Unprecedented transparency:** Required to upload since 2014
- ▶ **Why?**
 - ▶ Local court oversight

29 million Chinese court cases



- ▶ **Unprecedented transparency:** Required to upload since 2014
- ▶ **Why?**
 - ▶ Local court oversight
 - ▶ Commercialization of data related to courts

29 million Chinese court cases



- ▶ **Unprecedented transparency:** Required to upload since 2014
- ▶ **Why?**
 - ▶ Local court oversight
 - ▶ Commercialization of data related to courts
 - ▶ Move toward court automation

29 million Chinese court cases



- ▶ **Unprecedented transparency:** Required to upload since 2014
- ▶ **Why?**
 - ▶ Local court oversight
 - ▶ Commercialization of data related to courts
 - ▶ Move toward court automation
 - ▶ Increase court power?

29 million Chinese court cases



- ▶ **Unprecedented transparency:** Required to upload since 2014
- ▶ **Why?**
 - ▶ Local court oversight
 - ▶ Commercialization of data related to courts
 - ▶ Move toward court automation
 - ▶ Increase court power?
- ▶ **HUGE POTENTIAL** for researchers

29 million Chinese court cases



- ▶ **Unprecedented transparency:** Required to upload since 2014
- ▶ Why?
 - ▶ Local court oversight
 - ▶ Commercialization of data related to courts
 - ▶ Move toward court automation
 - ▶ Increase court power?
- ▶ **HUGE POTENTIAL** for researchers
- ▶ But selective exclusion → what is **missing**?

29 million Chinese court cases



- ▶ **Unprecedented transparency:** Required to upload since 2014
- ▶ **Why?**
 - ▶ Local court oversight
 - ▶ Commercialization of data related to courts
 - ▶ Move toward court automation
 - ▶ Increase court power?
- ▶ **HUGE POTENTIAL** for researchers
- ▶ But selective exclusion → what is **missing**?
- ▶ **Today:** what is online? what might be missing? How can these be used?

To start: Data from Henan province

To start: Data from Henan province

- ▶ 1,058,986 cases scraped from provincial website

To start: Data from Henan province

- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?

To start: Data from Henan province

- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?
 - ▶ Early adopter of legal transparency

To start: Data from Henan province

- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?
 - ▶ Early adopter of legal transparency
 - ▶ Access to internal legal data allows for

To start: Data from Henan province

- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?
 - ▶ Early adopter of legal transparency
 - ▶ Access to internal legal data allows for
 - ▶ detailed comparison to official statistics

To start: Data from Henan province

- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?
 - ▶ Early adopter of legal transparency
 - ▶ Access to internal legal data allows for
 - ▶ detailed comparison to official statistics
 - ▶ interviews with officials uploading docs

To start: Data from Henan province

- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?
 - ▶ Early adopter of legal transparency
 - ▶ Access to internal legal data allows for
 - ▶ detailed comparison to official statistics
 - ▶ interviews with officials uploading docs
- ▶ Huge variation across courts in proportion of cases online

To start: Data from Henan province

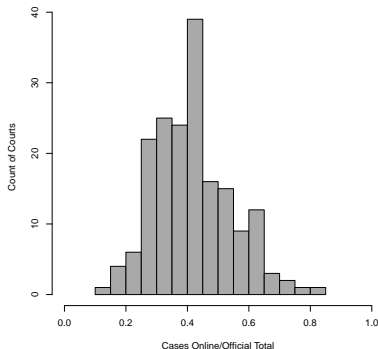
- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?
 - ▶ Early adopter of legal transparency
 - ▶ Access to internal legal data allows for
 - ▶ detailed comparison to official statistics
 - ▶ interviews with officials uploading docs
- ▶ Huge variation across courts in proportion of cases online

To start: Data from Henan province

- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?
 - ▶ Early adopter of legal transparency
 - ▶ Access to internal legal data allows for
 - ▶ detailed comparison to official statistics
 - ▶ interviews with officials uploading docs
- ▶ Huge variation across courts in proportion of cases online

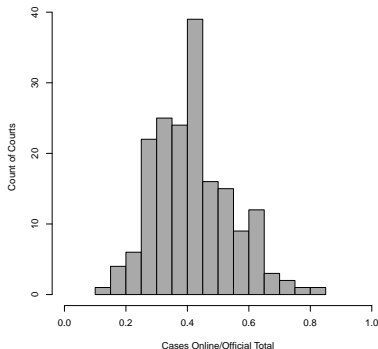
To start: Data from Henan province

- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?
 - ▶ Early adopter of legal transparency
 - ▶ Access to internal legal data allows for
 - ▶ detailed comparison to official statistics
 - ▶ interviews with officials uploading docs
- ▶ Huge variation across courts in proportion of cases online



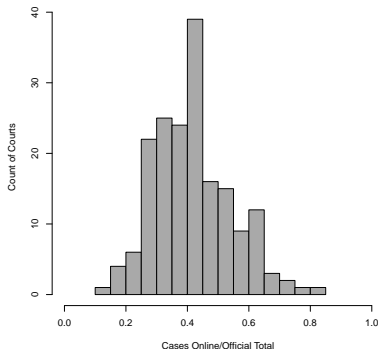
To start: Data from Henan province

- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?
 - ▶ Early adopter of legal transparency
 - ▶ Access to internal legal data allows for
 - ▶ detailed comparison to official statistics
 - ▶ interviews with officials uploading docs
- ▶ Huge variation across courts in proportion of cases online



To start: Data from Henan province

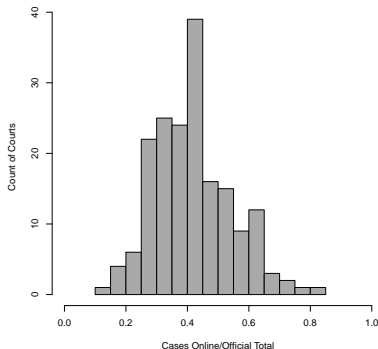
- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?
 - ▶ Early adopter of legal transparency
 - ▶ Access to internal legal data allows for
 - ▶ detailed comparison to official statistics
 - ▶ interviews with officials uploading docs
- ▶ Huge variation across courts in proportion of cases online



- ▶ intermediate courts more transparent than basic courts

To start: Data from Henan province

- ▶ 1,058,986 cases scraped from provincial website
- ▶ Why Henan?
 - ▶ Early adopter of legal transparency
 - ▶ Access to internal legal data allows for
 - ▶ detailed comparison to official statistics
 - ▶ interviews with officials uploading docs
- ▶ Huge variation across courts in proportion of cases online



- ▶ intermediate courts more transparent than basic courts
- ▶ no relationship with GDP or population

How to move forward with incomplete data?

- ▶ **Humility** about results, bounding

How to move forward with incomplete data?

- ▶ Humility about results, bounding
- ▶ Combine with qualitative work

How to move forward with incomplete data?

- ▶ **Humility** about results, bounding
- ▶ Combine with **qualitative** work
- ▶ Wait for data to get better

How to move forward with incomplete data?

- ▶ **Humility** about results, bounding
- ▶ Combine with **qualitative** work
- ▶ Wait for data to get better
- ▶ Work with **pockets** of good data (*)

Focus on a workable subset of the data

- ▶ From all types of tribunals, choose **administrative** tribunals

Focus on a workable subset of the data

- ▶ From all types of tribunals, choose **administrative** tribunals
- ▶ Then, choose tribunals with **> 50% transparency** and **> 20 cases** from 2013-2015

Focus on a workable subset of the data

- ▶ From all types of tribunals, choose **administrative** tribunals
- ▶ Then, choose tribunals with **> 50% transparency** and **> 20 cases** from 2013-2015
- ▶ Then, choose only **first instances** (not appeals, for instance)

Focus on a workable subset of the data

- ▶ From all types of tribunals, choose **administrative** tribunals
- ▶ Then, choose tribunals with **> 50% transparency** and **> 20 cases** from 2013-2015
- ▶ Then, choose only **first instances** (not appeals, for instance)
- ▶ We end up with 110 tribunal-years, 5200 decisions

Focus on a particular problem

- ▶ Which types of cases take the **longest** to resolve?

Focus on a particular problem

- ▶ Which types of cases take the **longest** to resolve?
 - ▶ legally and/or politically complicated?

Focus on a particular problem

- ▶ Which types of cases take the **longest** to resolve?
 - ▶ legally and/or politically complicated?
 - ▶ more difficult for the court?

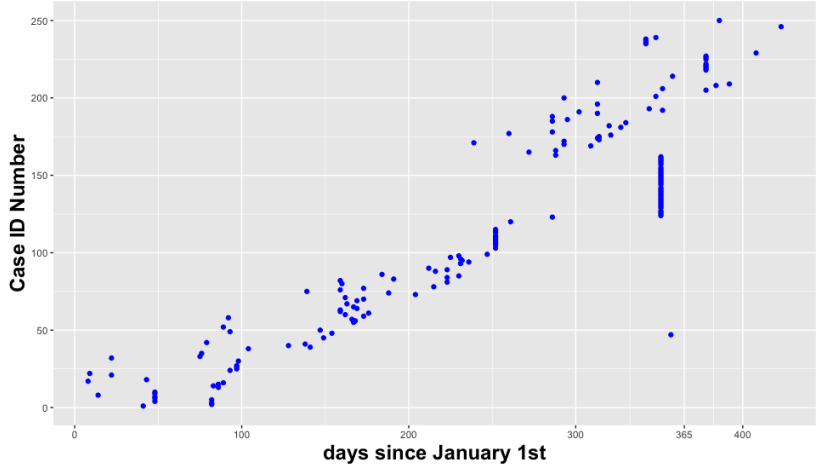
Focus on a particular problem

- ▶ Which types of cases take the **longest** to resolve?
 - ▶ legally and/or politically complicated?
 - ▶ more difficult for the court?
- ▶ First, estimate **start date** of case using Case ID number

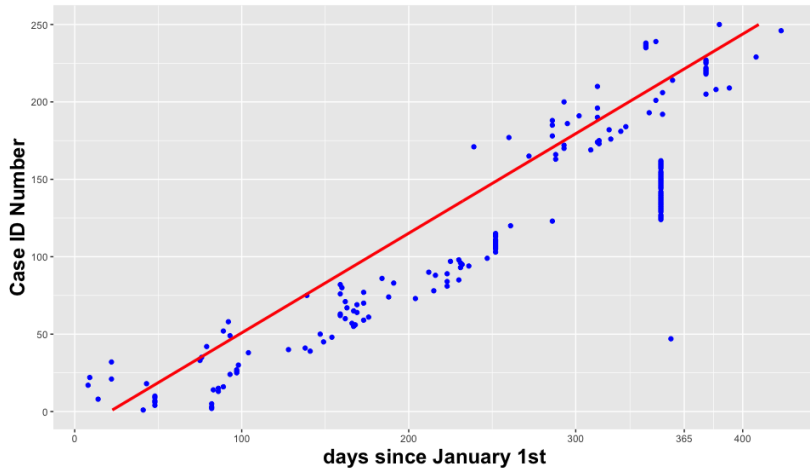
Focus on a particular problem

- ▶ Which types of cases take the **longest** to resolve?
 - ▶ legally and/or politically complicated?
 - ▶ more difficult for the court?
- ▶ First, estimate **start date** of case using Case ID number
- ▶ Decision date - start date = **length to decision**

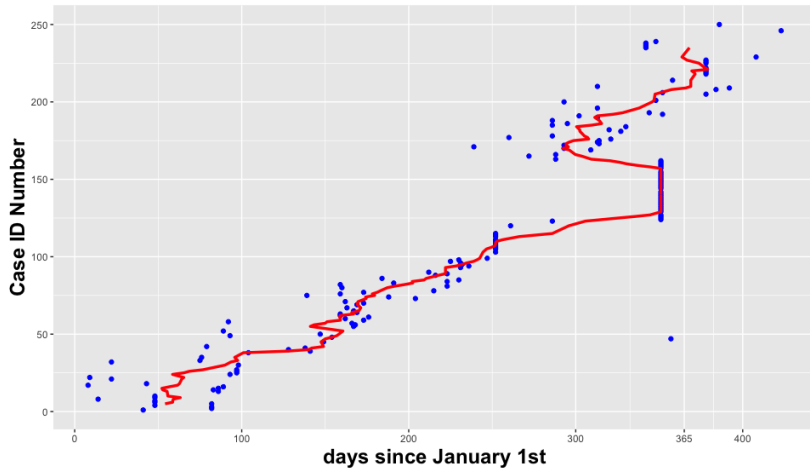
Dates and Case IDs



Method 1: Estimate start date



Method 2: moving average



Topics and time to decision

- ▶ Text as treatment method from Fong and Grimmer (2016)

Topics and time to decision

- ▶ Text as treatment method from Fong and Grimmer (2016)
 - ▶ Input text and time to decision

Topics and time to decision

- ▶ Text as treatment method from Fong and Grimmer (2016)
 - ▶ Input text and time to decision
 - ▶ Retrieves topics related to shorter and longer decisions

Topics and time to decision

- ▶ Text as treatment method from Fong and Grimmer (2016)
 - ▶ Input text and time to decision
 - ▶ Retrieves topics related to shorter and longer decisions
- ▶ Very much in exploratory stages in our training set

Preliminary results

- ▶ Cases about land use, permits and planning take a long time

Preliminary results

- ▶ Cases about land use, permits and planning take a long time
 - ▶ Permits and land use are longest

Preliminary results

- ▶ Cases about land use, permits and planning take a long time
 - ▶ Permits and land use are longest
 - ▶ Compensation takes shorter, but varies

Preliminary results

- ▶ Cases about land use, permits and planning take a long time
 - ▶ Permits and land use are longest
 - ▶ Compensation takes shorter, but varies
- ▶ Cases about workplace injuries take a short time

Preliminary results

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
	method award third countryside standard distribute land mu relationship policy	county funds five town receive country arable land department Petition grant	Planning Bureau planning project permit construction license city and countryside position issued influences	opinion construction planning land project main body evidence in summary before license	county funds receive five qualified immediately town amount farmer punish	finds work injury injure relationship human capital accident clear the reason the burden of proof fact
Time	-0.017	0.235	0.345	0.709	-0.523	-0.236