Bonustrack: Applications Using Online Data

Michael Kummer

NovaSBE

April 29, 2024

E-Commerce

Price comparison sites

Definition

- Website to compare prices for the same product at different shops.
- Examples are *shopper.com*, *market.yandex.ru* or *www.geizhals.at*.
- Consumers can compare prices and easily find the cheapest offer.

How can shops continue to make a profit on such a site?

- Can shops exploit behavioral patterns reap small profits?
- Chapters 4 and 5 focus on geizhals.at and two specific patterns.
  - Market Structure and Market Performance over a product's life
  - Specific behavioural feature - 9ending prices.

Price search engine *www.geizhals.at*

Background about the site underlying the data

- market leader in Austrian online price-comparison
  - online since late 1990s
  - commercial site since 2000
  - expansion to Germany, Poland and UK
- data are available since 2006
- daily prices for about 100,000 products and roughly 700 firms
- demand measured as clicks on e-tailers' homepage or last-click-through

www.Geizhals.at Search Result 1

www.Geizhals.at Search Result 2

Preis in €*	Anbieter	Händler-Bewertung	Verfügbarkeit / Versand*	Artikelbezeichnung des Händlers
311,--	T-Online-Shop [zum Shop] Hinweis: Firmensitz in Deutschland Infos AGB Meinungen	Note: 2,02 214 Bewertungen	versandfertig in 3 Tagen Vorkasse € 9,90 Nachnahme € 15,90 Kreditkarte € 9,90	Nokia E 61 Blackberry Silber - Mobiltelefon ohne Vertrag ! (23.11.2007, 10:11)
336,--	it-designworks (haym.infotec) Infos AGB Meinungen	Note: 1,20 850 Bewertungen	versandbereit in 1-2 Werktagen Vorkasse € 5,88 Nachnahme € 9,38 Express ab € 13,80 Abholung möglich (A-5550 Radstadt)	Nokia 0040446 Nokia - E61 UMTS-Mobiltelefon grau (E61) (0040446) (Art# 7A31909) (23.11.2007, 10:13)
348,24	MTS-Shop Infos AGB Meinungen	Note: 1,38 200 Bewertungen	Lagernd im Außenlager, 1-2 Werktage Abhol/Versandbereit (Ab Bestellbestätigung) Österreich: Vorkasse € 5,- Nachnahme € 11,- Kreditkarte € 18,93 Express € 25,50 Deutschland: Vorkasse € 13,40 Nachnahme € 26,40 Kreditkarte € 27,33 Abholung nach Vereinbarung möglich (A-1170 Wien)	Nokia 05no0061 Nokia E61 grey/silver (E-)GPRS, HSCSD • WAP/MMS • Push-to-talk • OS: Symbian Series 60 (3rd Edition) • Quac Mio. Farben, 320x240 Pixel) • Vibracall • polyphone Klingeltöne • Video-/Mp3-Player • J.. (23.11.2007, 10:15)
363,--	TC4U Infos AGB Meinungen	jetzt bewerten!	2-4Werktage Österreich: Vorkasse € 5,- Nachnahme € 9,- Deutschland: Vorkasse € 10,-	Nokia E61 (23.11.2007, 10:11)
364,95	SYShop Infos AGB Meinungen	Note: 1,70 22 Bewertungen	**Versandbereit in 2 - 4 Werktagen** **Stand: 23.11.2007, 14:56 Uhr** Österreich: Vorkasse € 4,80 Nachnahme € 9,90 Deutschland:	Nokia Nokia E61 ohne Bindung Nokia E61 grey/silver (E)GPRS/HSCSD (A) WAP/MMS Quadband RS-MMC-Slot Video/Mp3 144g (23.11.2007, 10:11)

- Austria's largest price-comparison site.
- Users search for specific items.
- Lowest price listed on top.
- Shipping cost posted right beside.
- More characteristics (reputation, availability, ...).

Paper 1: 99 Cent: Price Points in E-Commerce

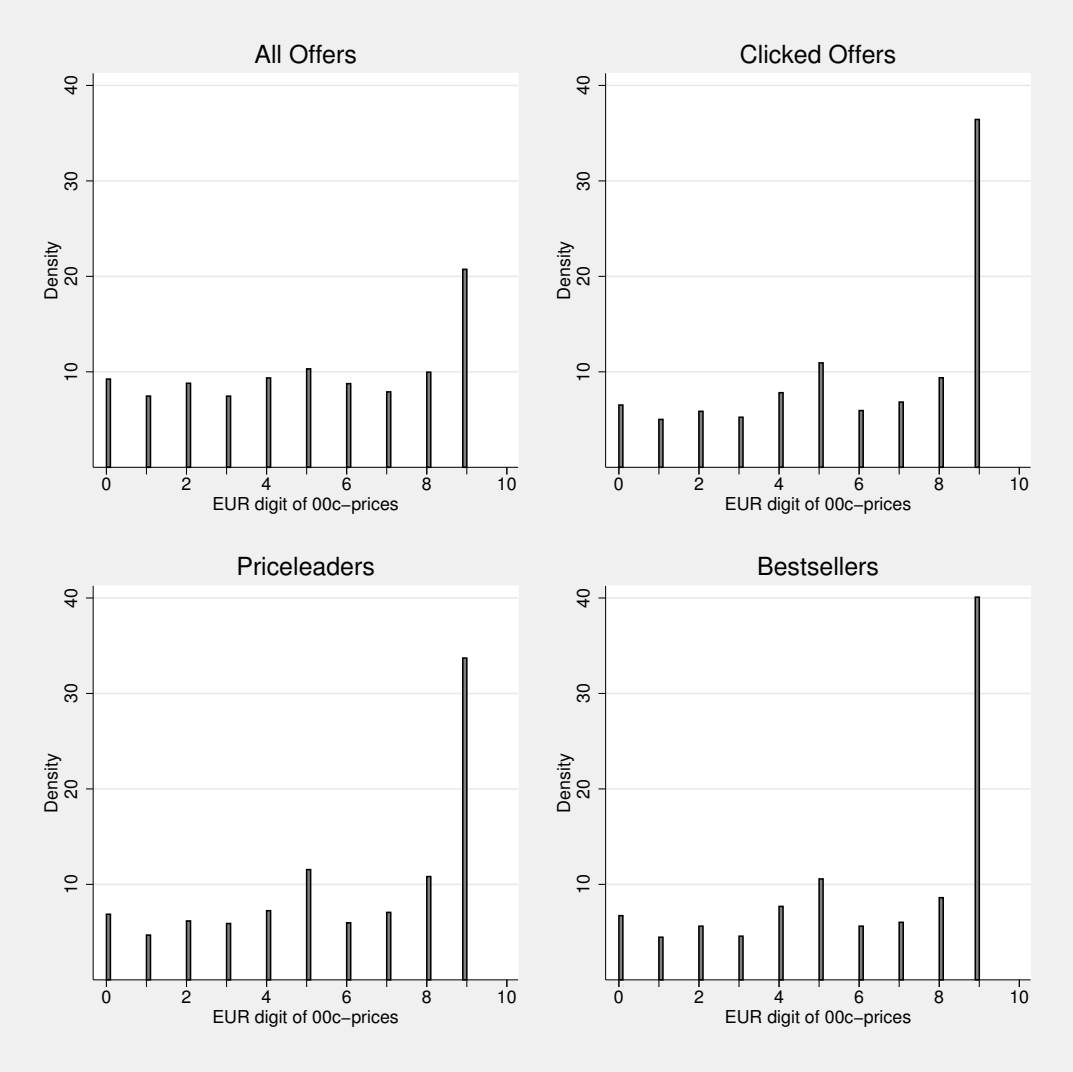
- Why Do Firms Use 99 Cent Prices? Is there a level effect?
  - Odd prices: 99 Cents and variations
  - Even prices: zero-ending
- We test demand effects and price stickiness for price points online.
  - Prediction 1: demand reacts only to Euro differences.
  - Prediction 2: .99 prices more sticky.

Dataset:

- new price quotes in one week in June 2007 (inflow sample)
- 810,000 prices, 30,000 products, 700 firms
- use only products with price larger than €25.

Distribution of last Euro digits of 00c-endings.

- 9.00 is twice as frequent as other even Euro-prices
- Among clicked and widely sold products, 9.00 constitutes about 33% of even prices.
- Same pattern for special endings (00c, 50c, 90c & 99c)



Conclusions

Conclusions I: Focal Prices are more sticky.

- Duration Analysis for survival of focal (9-ending) prices.
  - firms maintain 99 cent prices longer (also longer than 00 prices)
  - 9-endings are not so easily undercut if they are price-leaders

Conclusions

Conclusions I: Focal Prices are more sticky.

- Duration Analysis for survival of focal (9-ending) prices.
  - firms maintain 99 cent prices longer (also longer than 00 prices)
  - 9-endings are not so easily undercut if they are price-leaders

Conclusions II: Focal Prices do not have lower demand (not shown)

- consumers do disregard cents after the euro

Paper 2:

Market Structure and Market Performance in E-Commerce

Market Structure and Market Performance in E-Commerce

- **Question:** How do more sellers affect Markup & Price Dispersion?
  - Study the relationship of market structure and market performance in e-commerce.
  - Does the relation between # of firms and markup change over the product lifecycle?
- **Worry:** Correlation of Number of Firms and Markups mismeasured.
  - endogeneity?
  - products with higher markups attract more sellers!
- **Solution:** Exploit Listing Patterns across Lifecycles.
  - shops that are always first to list a product should be early on any new product.
- **Contribution:** Instrumental Variable and Application.
  - Valid proxy for markups, because we can measure wholesale prices
  - Propose novel instrument for number of firms in online markets
  - Analysis of Market Structure and Performance over the Products' Lifespan

Dataset and organisation of data

2 Data Sources

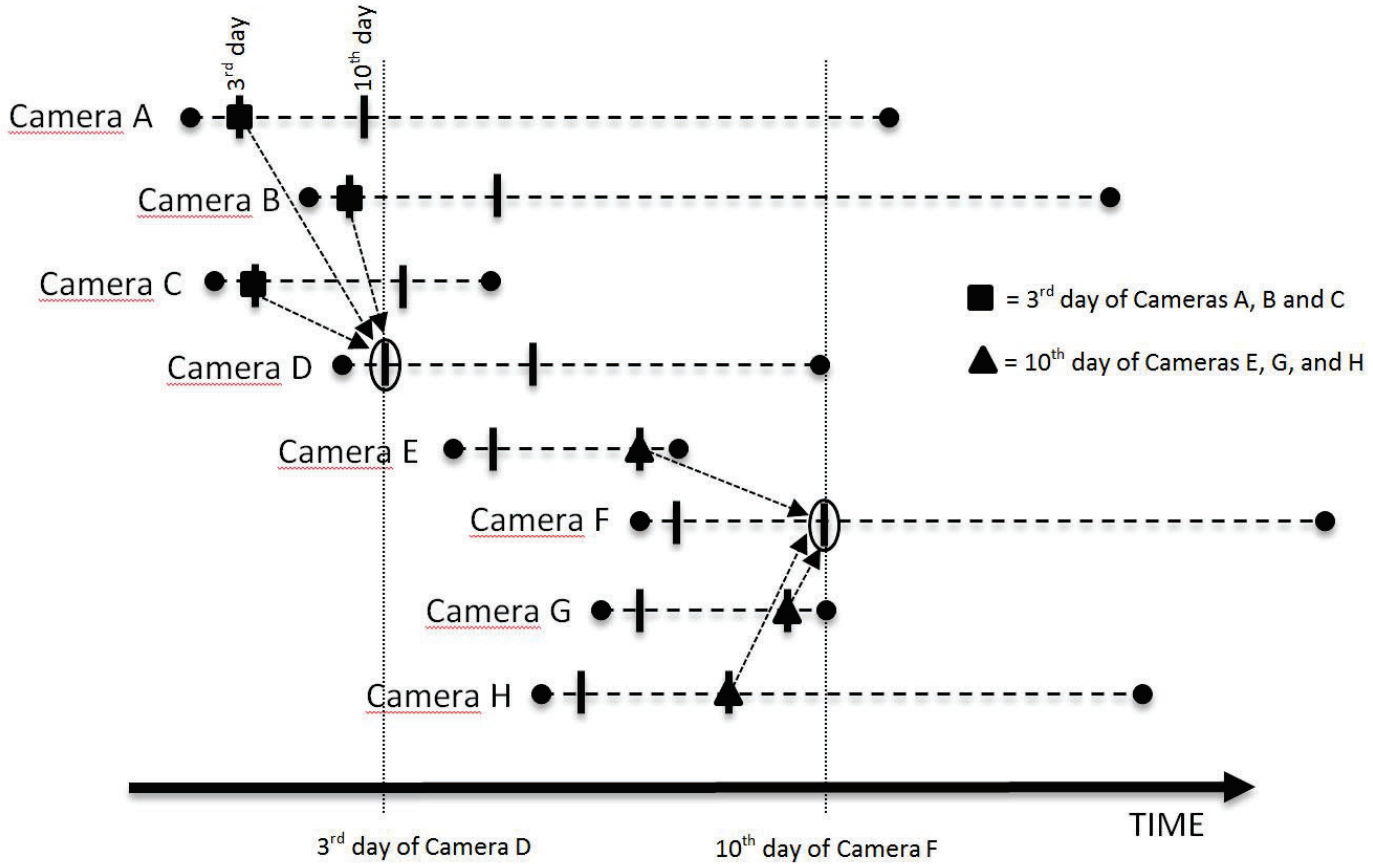
- Augmented geizhals-database with Information on wholesale-prices from a major producer of digital cameras, scanners etc.

Organization:

- 70 products
- Unit of observations is market by day: 15893 observations.
- Computed minimum and median markup and price dispersion
- re-aligning all beginning dates

$=>$ Study product life cycle, observe all products from birth to death

Instrumentation Strategy



- use firm's listing behavior in earlier lifecycles
- for a fixed number of products, compute average number of listing firms
- use only the same specific lifecycle day of the earlier products
- very strong instrument (predictor is highly significant, F-Test-Statistic well above 10)

Conclusions

1. Market Structure and Performance of E-commerce
   - Number of sellers has a negative effect on the median price.
   - Results confirmed for markup of the price-leader (minimum markup):
     - ★ decreases by 4.5% if number of sellers increases (one std. deviation)
   - Results for price dispersion do not allow for strong conclusions
     - ★ weak evidence in favor of search theoretic models.

2. Further results: Product Life Cycle Effects
   - Relationship remains largely stable over the lifecycle.
   - Competition-effects differ for same-brand and competitor's products.

Key Challenges

- Dump is huge (Runtimes)
- Data Cleaning in the Beginning
- SQL

WIKIPEDIA

Spillovers in Content Networks -

→ Focus on the hyperlinked citation network between Wikipedia's articles

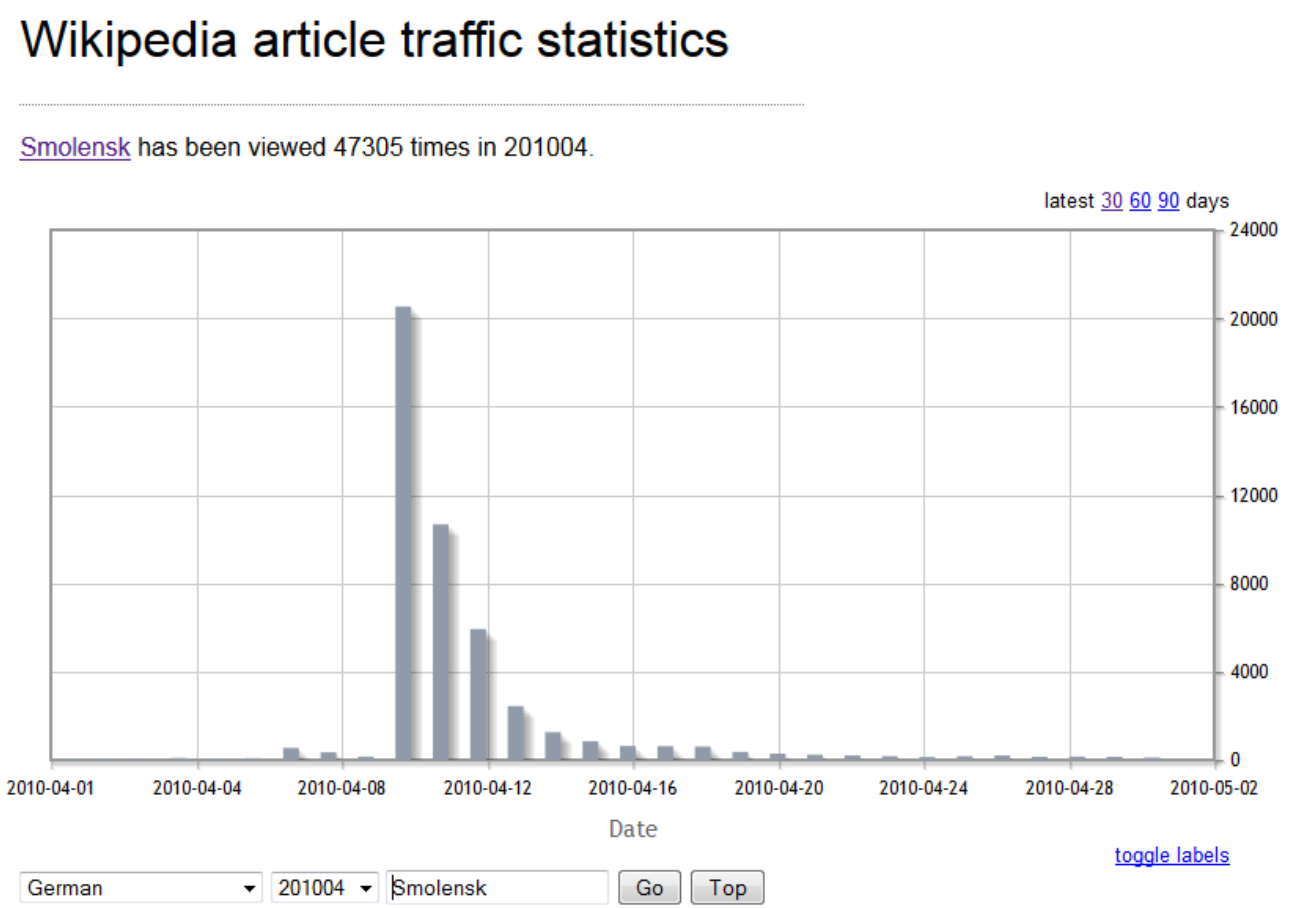
- **Question:** How do citations influence user search and contributions?
  - How much attention spills from one article to the next?
  - Does attention drive contributions? Which articles will evolve?
  - Relates to production in other "citation networks" (Open Source, science)
  - Also relates to questions of advertisement.
- **Worry:** Correlations of Network and Outcome not causal.
  - endogeneity?
  - unobserved factors drive both?
- **Solution:** Exogenous local Shocks to single Nodes.
  - activity increase on neighbors, due to link from treated page
  - pseudo-experimental design
- **Contribution:** Formal Framework and Application.
  - extension of Bramoullé et al. (2009) to include exogeneous shocks.
  - analysis of 23 disasters and 34 "Today's featured articles" on Wikipedia

Wikipedia

Background and Data

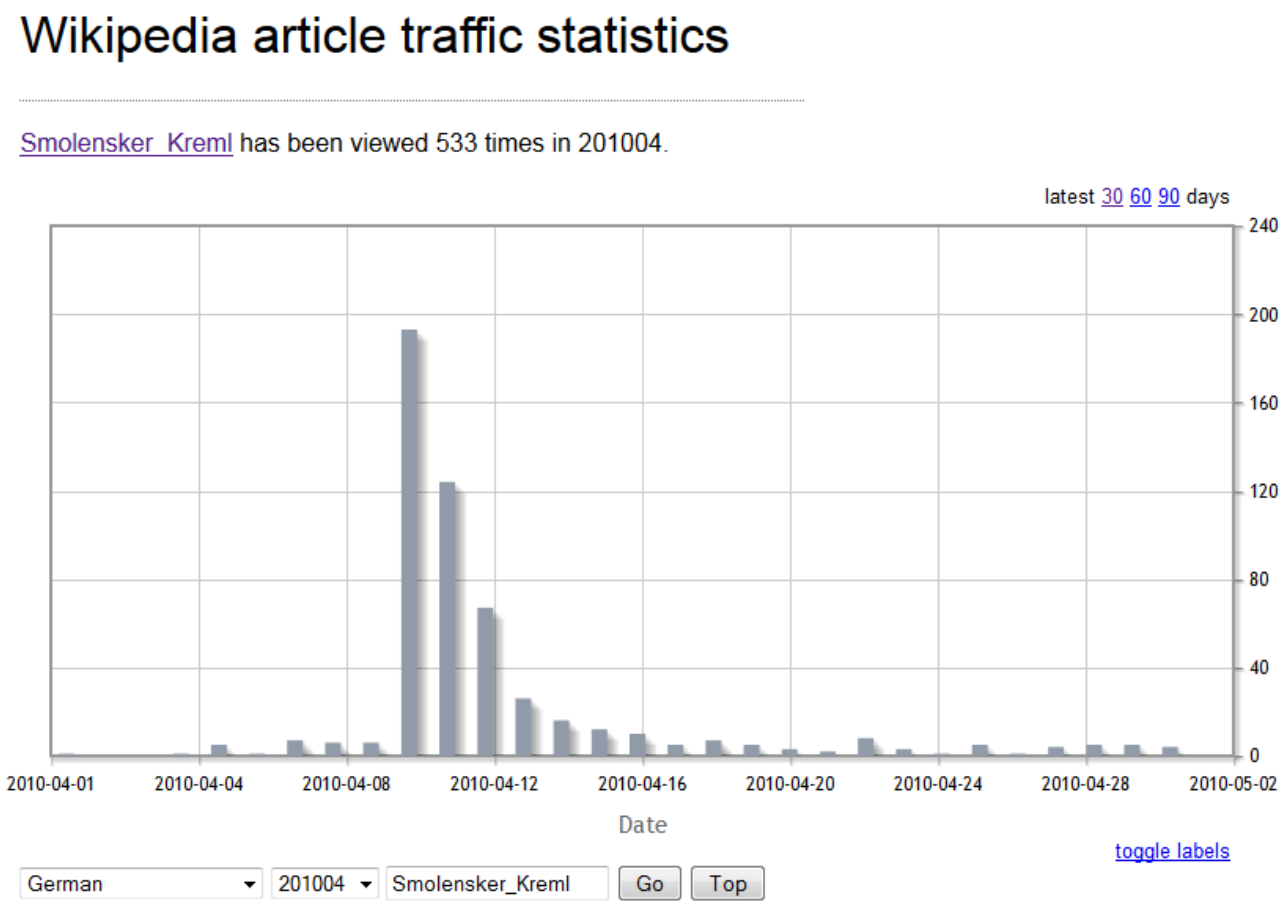
- World's largest platform for encyclop. knowledge.
- >200 languages, >54 with >100,000 Articles. 1,716,000 in German
- 1000s of volunteers
- public metatext-dump augmented with clicks (Dec. 2007 - Dec 2010.) and articles-links.
- 29 days of data for 57 special articles and neighbors.

Main Idea: Illustrated by the case of Smolensk...



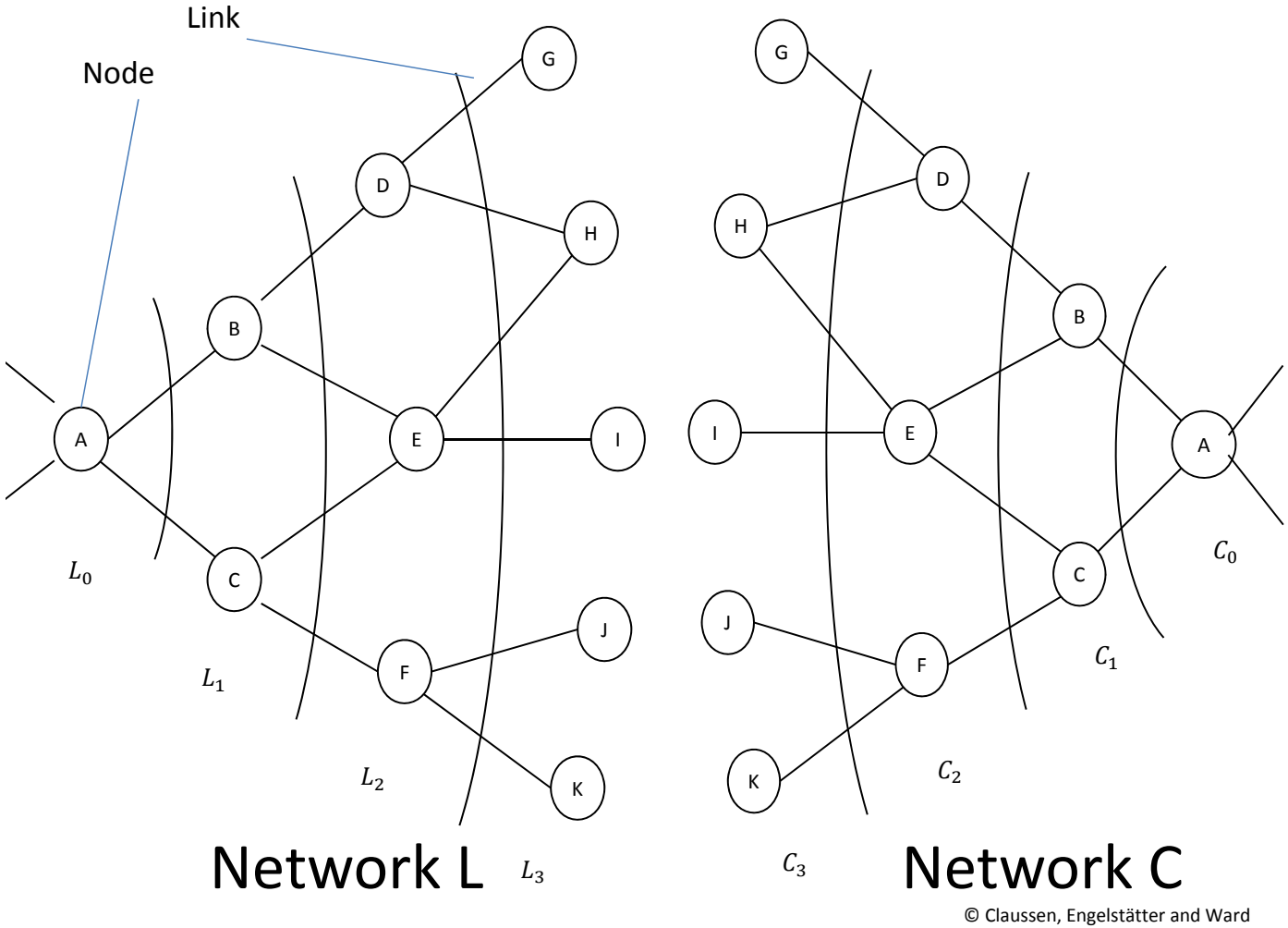
On April 10th, 2010 the plane of the Polish government crashed

Example: 1 click from Smolensk (Smolensk Castle)

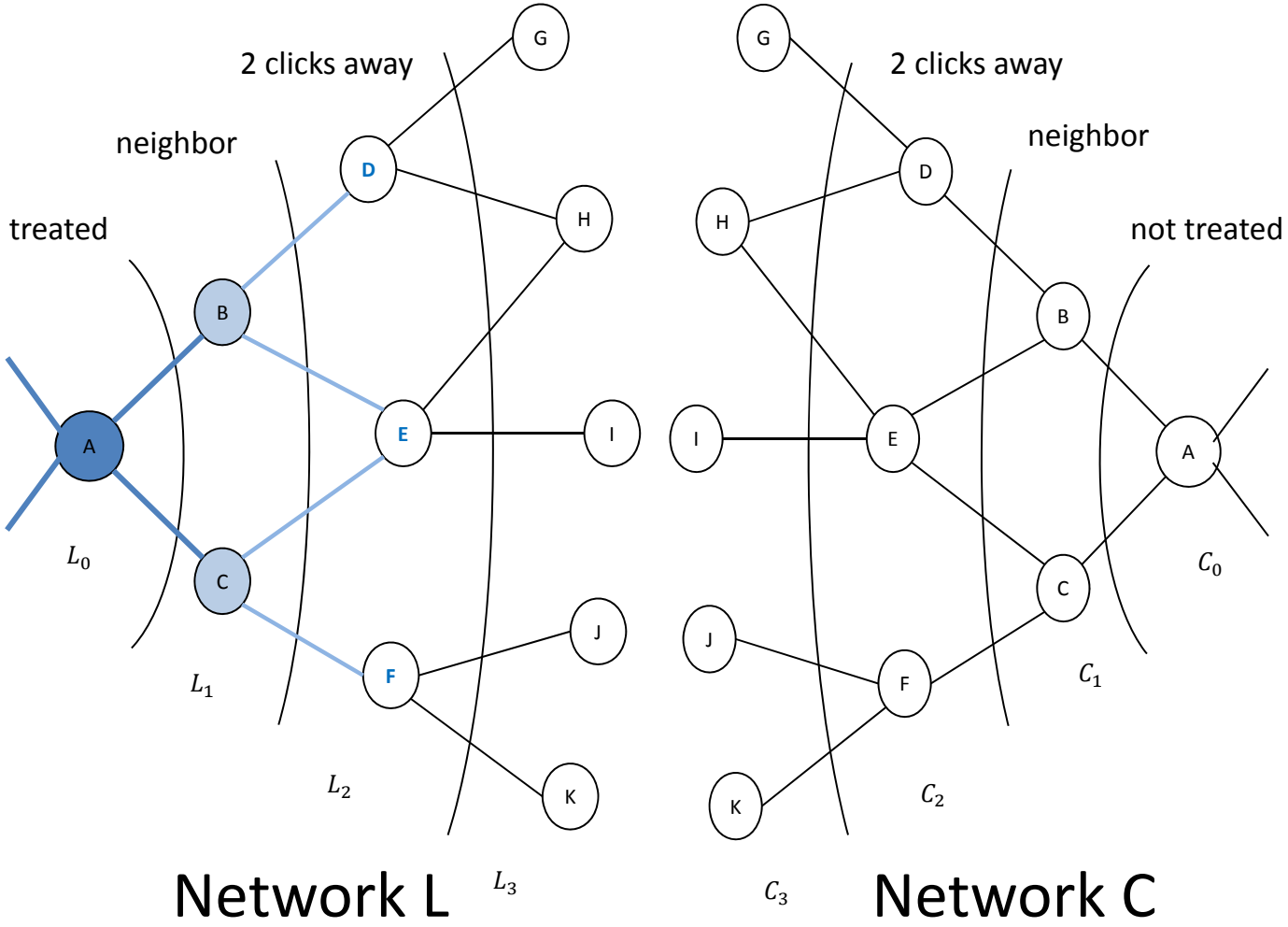


On April 10th, 2010 the plane of the Polish government crashed
The Castle of Smolensk was not directly affected

General Idea: Network a Pond with two Groups



A Stone in a Pond: Treat only one



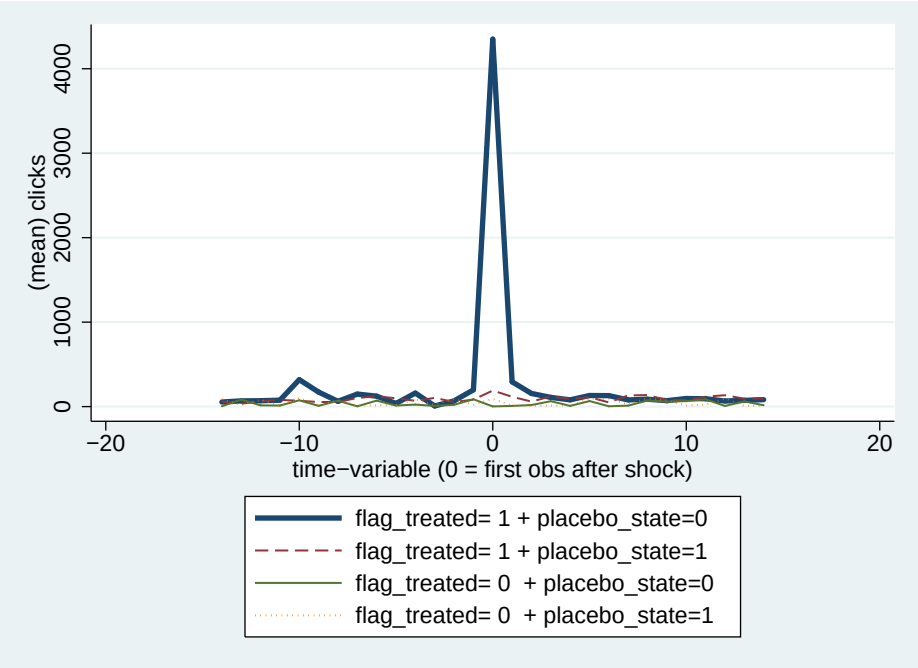
Two Types of Shocks

1. Large Media Event: Sichuan Earthquake
   - Date: 12.5.2008
2. Today's "featured article" ("Seite des Tages"): Banjo-Kazooie
   - featured on the 10.06.2010

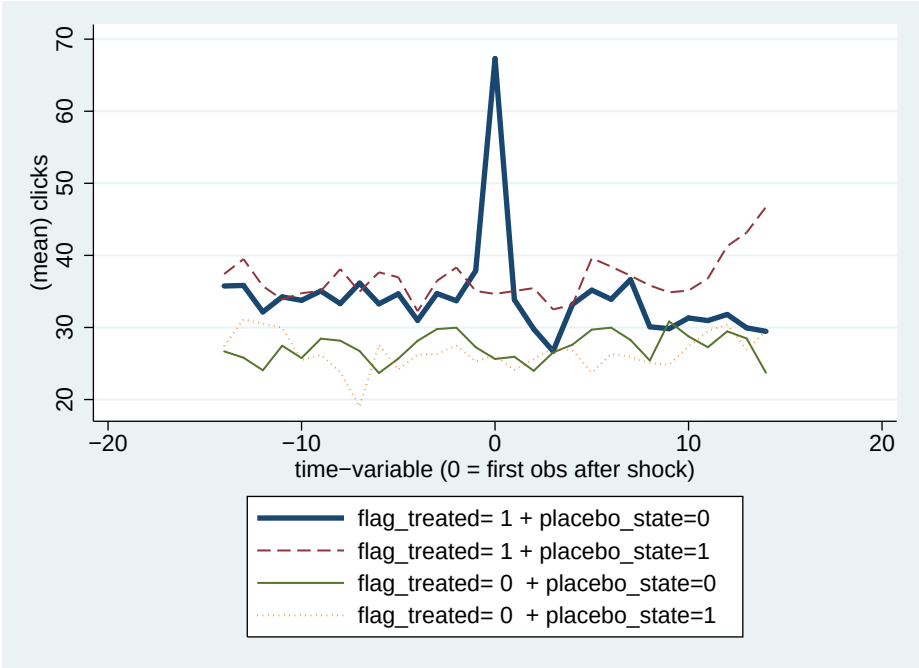
Application

Clicks from advertised articles to neighbors

4,200 more clicks on featured page.



40 more clicks on neighbors.



- 36 “Seite des Tages” - experiments (and 36 controls)
- Number of neighbor pages: approx. 6600
- 4000 clicks on L0 $\rightarrow$ 4000 clicks and 5 edits on the neighbors (not shown)

Key Challenges

- Dump is huge (Runtimes)
- Data Cleaning in the Beginning
- SQL

Wikipedia matters!

Does Wikipedia Content matter?

Main Ideas

- **Question:** Does information in Wikipedia affect tourists' choices?

Does Wikipedia Content matter?

Main Ideas

- **Question:** Does information in Wikipedia affect tourists' choices?
- **Method:** Experiment on Wikipedia with pages about Spanish cities!

Does Wikipedia Content matter?

Main Ideas

- **Question:** Does information in Wikipedia affect tourists' choices?
- **Method:** Experiment on Wikipedia with pages about Spanish cities!
  - Add about 2 paragraphs of text and a photo

Does Wikipedia Content matter?

Main Ideas

- **Question:** Does information in Wikipedia affect tourists' choices?
- **Method:** Experiment on Wikipedia with pages about Spanish cities!
  - Add about 2 paragraphs of text and a photo
  - Track how many tourists decide to visit.

Does Wikipedia Content matter?

Main Ideas

- **Question:** Does information in Wikipedia affect tourists' choices?
- **Method:** Experiment on Wikipedia with pages about Spanish cities!
  - Add about 2 paragraphs of text and a photo
  - Track how many tourists decide to visit.
- **Result:** our treatment increases hotel stays by 9%

Does Wikipedia Content matter?

Main Ideas

- **Question:** Does information in Wikipedia affect tourists' choices?
- **Method:** Experiment on Wikipedia with pages about Spanish cities!
  - Add about 2 paragraphs of text and a photo
  - Track how many tourists decide to visit.
- **Result:** our treatment increases hotel stays by 9%
  - Result is driven by shorter Wikipedia pages (pages where our treatment was relatively larger)

Data sources

1. Wikipedia dataset
   - Characteristics of Wikipedia articles of Spanish cities
   - In different language Wikipedias (in languages of tourist origin)
   - For each article, the length of text & number of photos (over time), survival of text and photos
2. Tourism data from the Spanish National Institute of Statistics
   - Overnight hotel stays in Spanish cities
   - By the country of tourist origin
   - Focus on 4 countries of origin: Germany, France, Italy, the Netherlands
   - Years 2010–2015

- Wikipedia page views
  - Number of visits to Wikipedia articles by language
- Google Trends
  - Search volume for the Spanish cities by country

Experimental design: sample

- Starting point: sample of 135 Spanish cities
  - For which we had data on hotel stays
- Restricted attention to cities that satisfied 2 criteria:
  1. Wikipedia page relatively short
     - ★ In all 4 languages (Dutch, French, German, Italian), Wikipedia page no longer than 24,000 characters
  2. No missing data on hotel stays
     - ★ In case of tourists from all 4 countries, data on hotel stays exists in all months from May until October in 2013
- 60 cities satisfied these 2 criteria
- 240 Spanish city – language (country) pairs

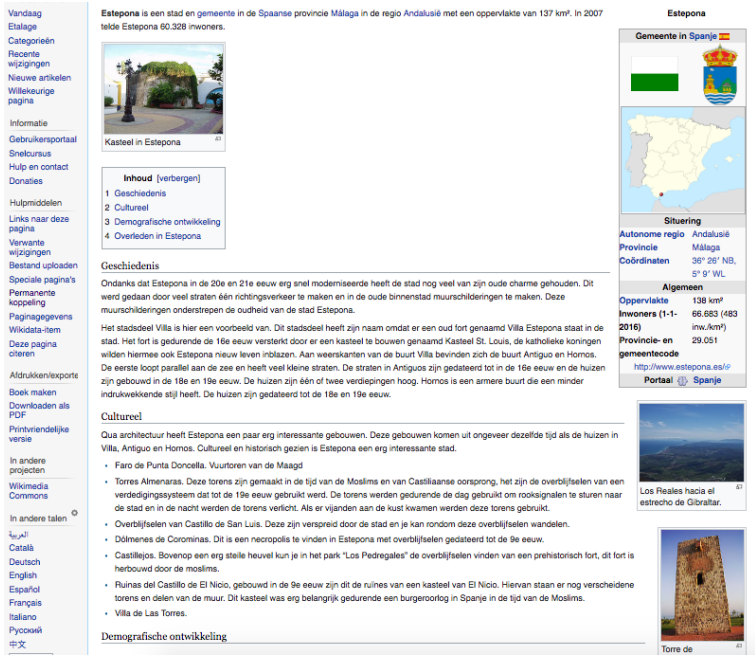
Example: Treated city pages in 2 random languages

Treatment:

Translate 2 paragraphs from the Spanish version, and add them to treated language (and a photo).

Survival of the text & photos we added to Wikipedia

(a) Both before & after



(b) Our treatment

- Dutch Wikipedia: all our additions were 100% deleted in 24h
- We exclude all Dutch articles from all our analysis
  - Results very similar if we consider Dutch articles as non-treated

Survival of the text & photos we added to Wikipedia

	France	Germany	Italy	Total
% text survived: 24h	100.0	94.7	100.0	98.2
% text survived: next month	98.7	90.2	99.9	96.3
% text survived: next year	95.1	86.7	97.5	93.1
% photos survived: 24h	100.0	96.2	100.0	98.8
% photos survived: next month	100.0	92.3	96.4	96.4
% photos survived: next year	100.0	88.5	92.9	94.0
Number of observations	30	30	30	90

- In German, French, & Italian our additions survived well
  - By the beginning of next month: on average 96%
  - By the beginning of next year: on average 93%

Comparison of treatment and control groups

Table: Dependent variable: treatment status

	Coef.	p-value
Log(Sum of tourists in 2013)	-0.002	0.958
Log(Number of tourists)	-0.012	0.527
Tourist data missing	0.045	0.556
Log(Initial text length)	-0.000	0.994
height		

- Each row presents estimates from a separate regression
- Rows 1 and 4: observation is a city-language pair
- Rows 2 and 3: observation is a city-language-month triplet, sample covers time period until treatment

Effect of the treatment on hotel stays

Table: Dependent variable: Logarithm (number of hotel nights)

	(1)	(2)
Treatment	0.089**	0.002
	(0.045)	(0.038)
Treatment: Small page		0.332***
		(0.100)
City-Language FE	Yes	Yes
Adj. R-squared	0.245	0.248
Observations	5688	5688
height		

- Observation: month-city-language (tourist country of origin pair)
- Sample: 2010–2015, May–October, tourists from Italy, France, Germany to the 60 cities in Spain
- Treatment: equals 1 for months after treatment for treated city-language pairs, and 0 otherwise
- Small page: Initial page size is below the 25th percentile
- Controls: indicator for period after treatment interacted with language FE, indicator for period after treatment interacted with city FE, logarithm of number of tourists from Spain interacted with language
- Standard errors clustered by city-language pair (180 clusters)

Robustness

- Sample: Included Dutch (assume Dutch pages not treated)
- Sample: Included all 12 months
- Sample: Included time periods with missing hotel data
- Control variable: Include logarithm of UK tourists
- Dependent variable: share of tourists from country x
  - instead of logarithm of the number of tourists from country x

- **Conclusion:** Wikipedia matters!
  - Implication: Wikipedia production & biases matter

Key Challenges

- Experiment was extremely costly (Money and Time)
- Dutch Community undid our Treatment.
- Had to wait for Spanish Data.
- Could not go back to improve our Design.

THANK YOU FOR YOUR ATTENTION!

THANK YOU FOR YOUR ATTENTION!

Michael Kummer
m.kummer@uea.ac.uk

Appendix

Additional Materials

GOOGLE PLAY STORE

Research Questions

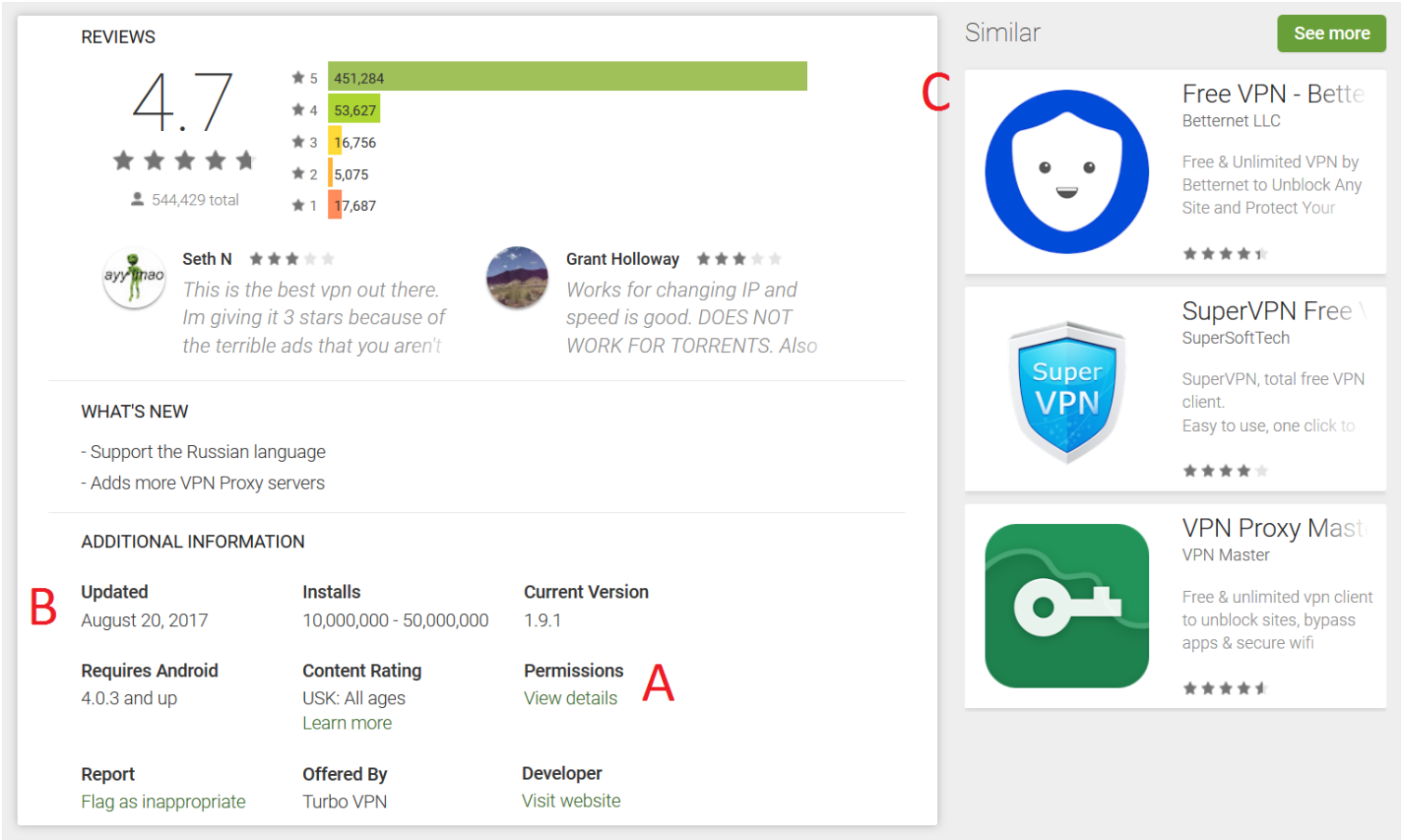
Question 1: Does **market concentration** (MC) affect **innovation?**

Question 2: Does **collection of user data** mediate this concentration-innovation relationship?

1. Does MC affect data collection of firms / privacy of users?
2. Does data collection affect innovation performance?

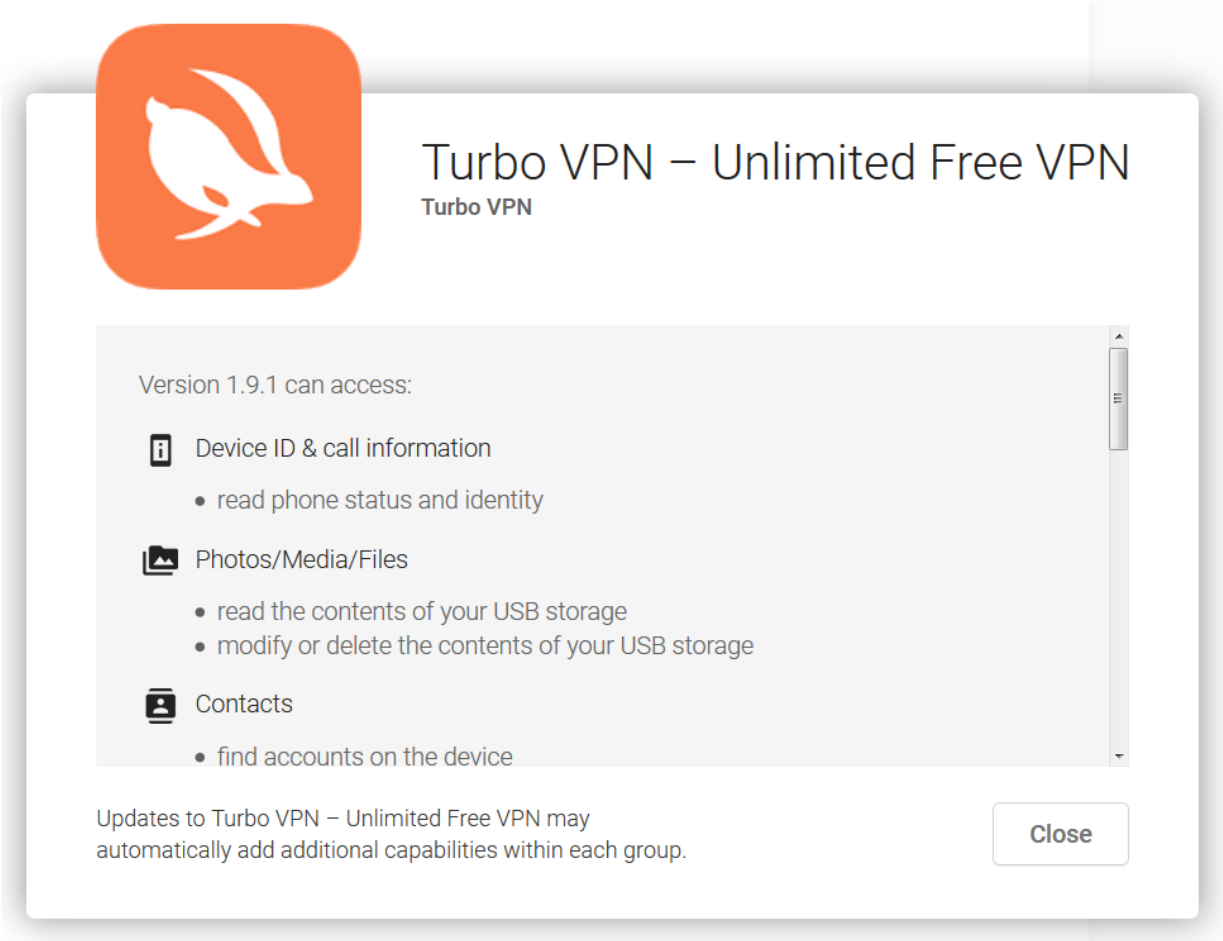
DATA

Data from Google Play Store



- We observe everything Play Store users can see about an app.
- Specifically here: (A) Privacy, (B) Updates, & (C) an app's Competitors.

Data A: Measuring Data Collection and Privacy

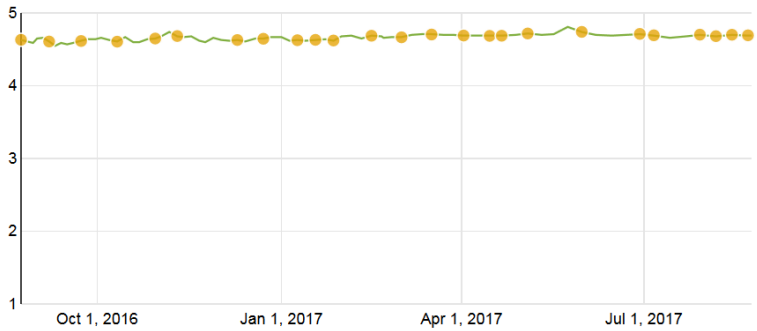


- On privacy: All permissions that apps request.
- may allow the app to track or identify the user, their contacts, etc...

Data B: Additional Information on Innovation and Privacy

10+ Million Installs	1 Million Recent installs	4.68 540,865 Rating	Top ranked Ranking	10 Libraries	4.0.3+ Android version	8/17/17 Last updated	2016 June App Age	6.89 MB App size	Everyone Content Rating	FREE Price

Google Play Rating history and histogram



Changelog

Aug 17, 2017	UPDATE	Version 1.9.1
Aug 5, 2017	UPDATE	Version 1.9.0
Jul 30, 2017	UPDATE	Version 1.8.9
Jul 25, 2017	UPDATE	Version 1.8.8
Jul 4, 2017	UPDATE	Version 1.8.6
Jun 22, 2017	UPDATE	Version 1.8.2
May 31, 2017	INSTALLS	10,000,000+ installs

View more

Ad network libraries

AdMob

Social libraries

facebook Facebook

Development tools

Android Support library

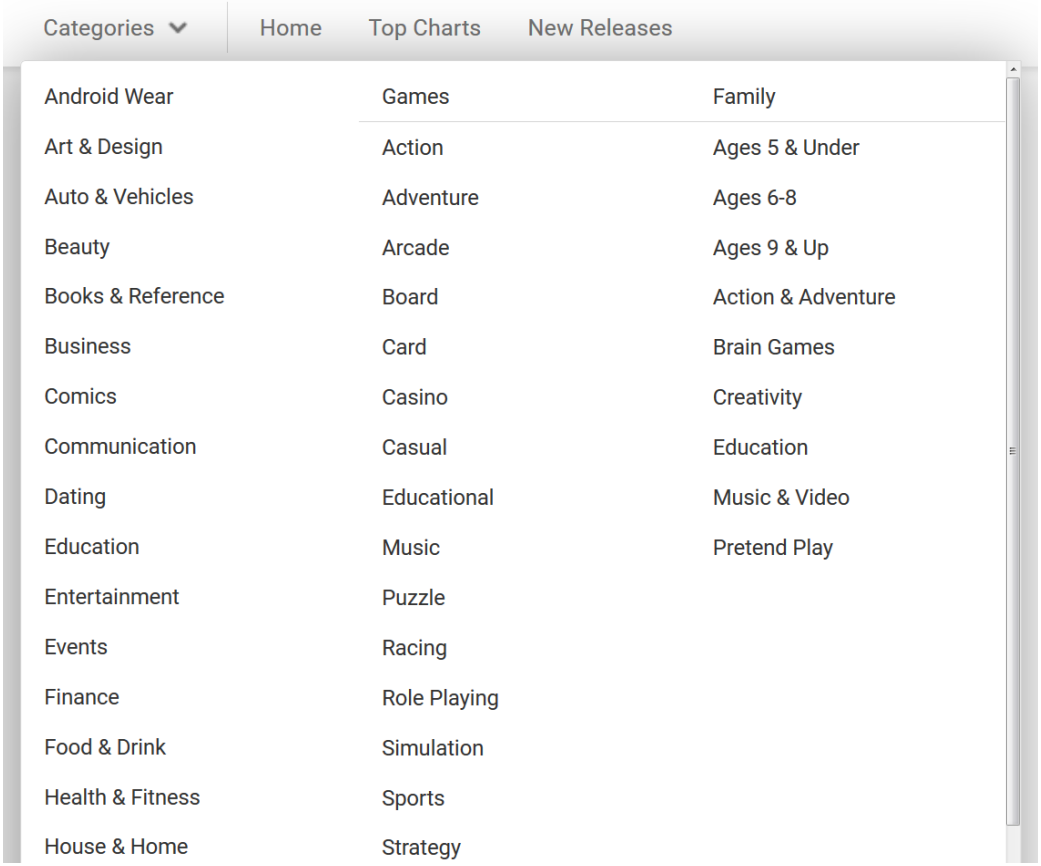
FLURRY Flurry Analytics

⋮

- We additionally observe the last 50 updates and logging information...
- ...and the third-party libraries that the app shares its data with.

Data C: Defining the Markets and Concentration I

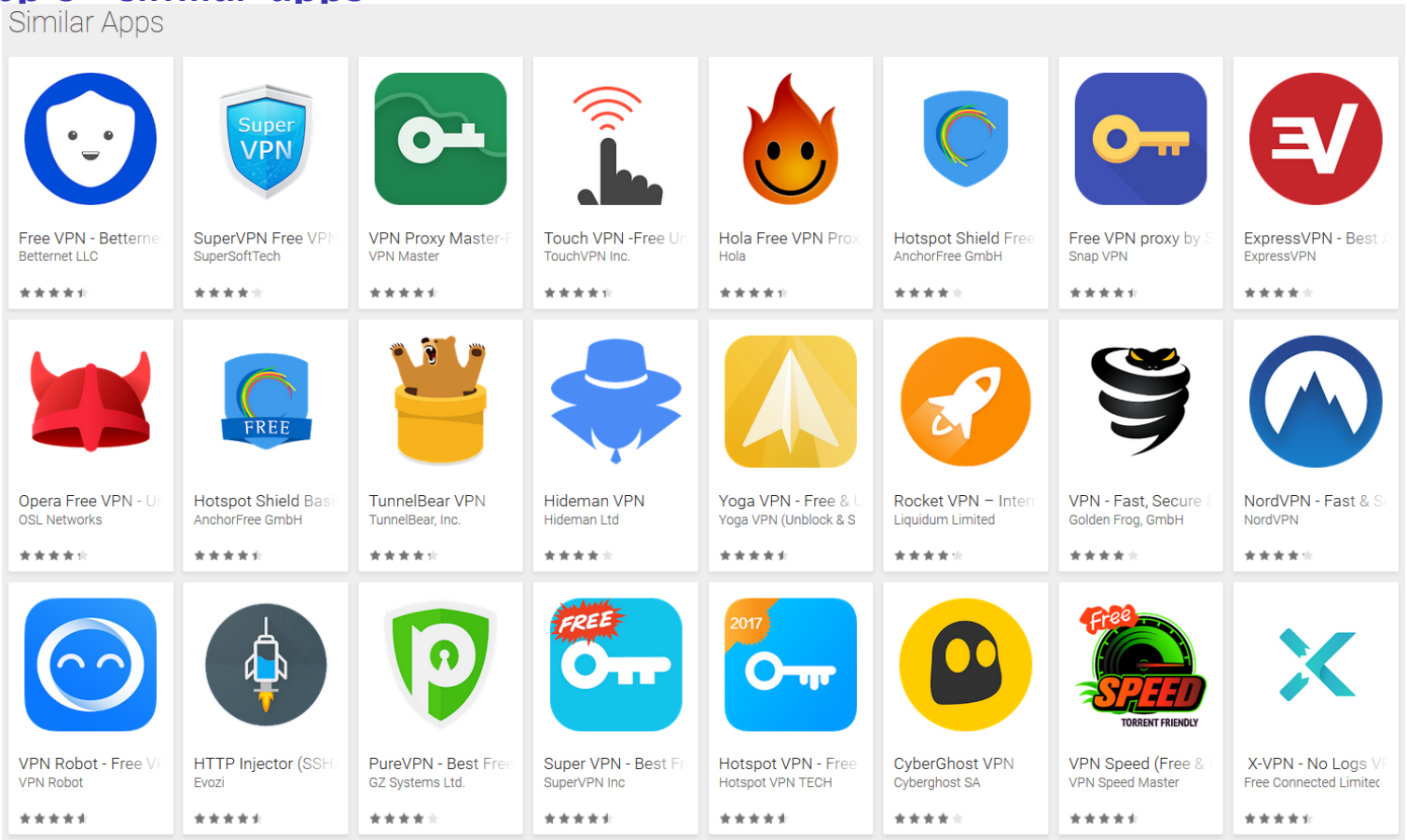
via categories



- The roughest definition of a market would be the Play Store's categories.
- But with 40 categories that is too rough. Instead...

Data C: Defining the Markets and Concentration II

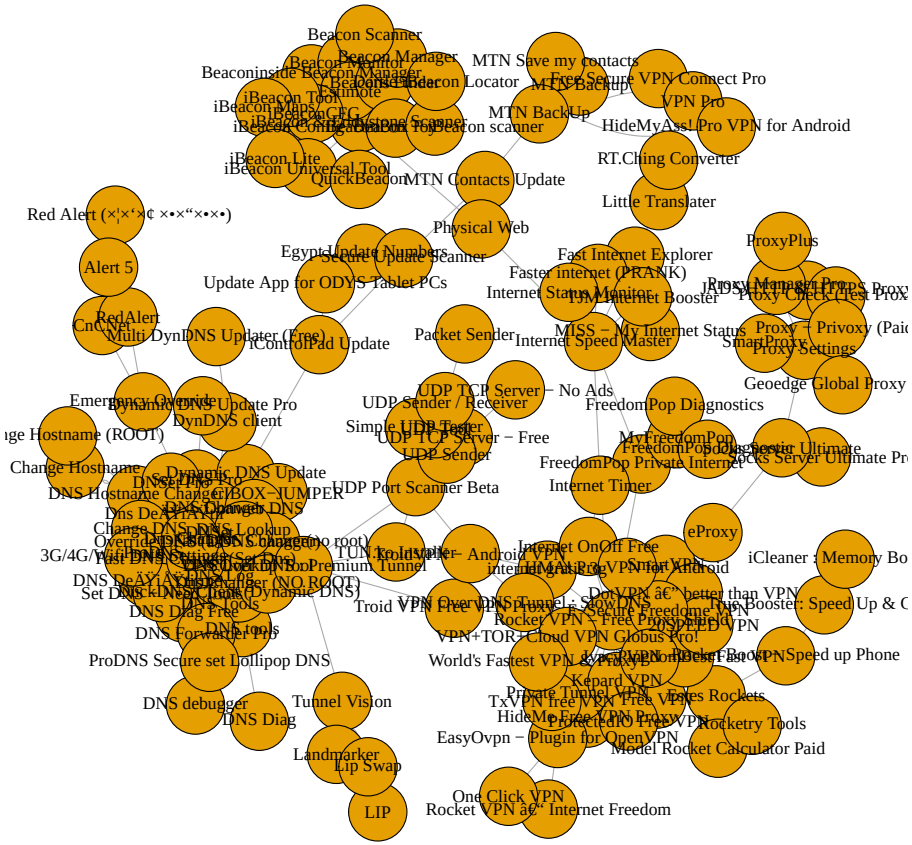
via an app's "similar apps"



- ...we define a market using all "similar apps" as competitors.
- This gives us a 'naive' app-specific market definition (CS1 & Panel1), or...

Data C: Defining the Markets and Concentration III

via the network of "similar apps"



- ...we find network clusters on the network formed by "similar apps".
- Each of our clusters is a market of its own (CS2 & Panel2).

Note: An app is a node, and a "similar app"-link is an edge here.

Data: Overview

Panel of nearly 2 million apps in Google's Play Store; quarterly for 2015-2017

A: Data Collection

- Measured via privacy-sensitive permissions ($D_{DataCollection}$ and $\#_{DataCollection}$)

B: Innovation

- Major update ($MajorUpdate_{t+1}$): has a large change in either (a) app description, (b) version number or (c) number of unproblematic permissions

C: Market Concentration

- Competitors identified by (1) "'similar apps" (2) and app clusters in network
- Market share based on number of new ratings ($\frac{\Delta Ratings}{\sum \Delta Ratings}$)
- Market concentration measured by HHI

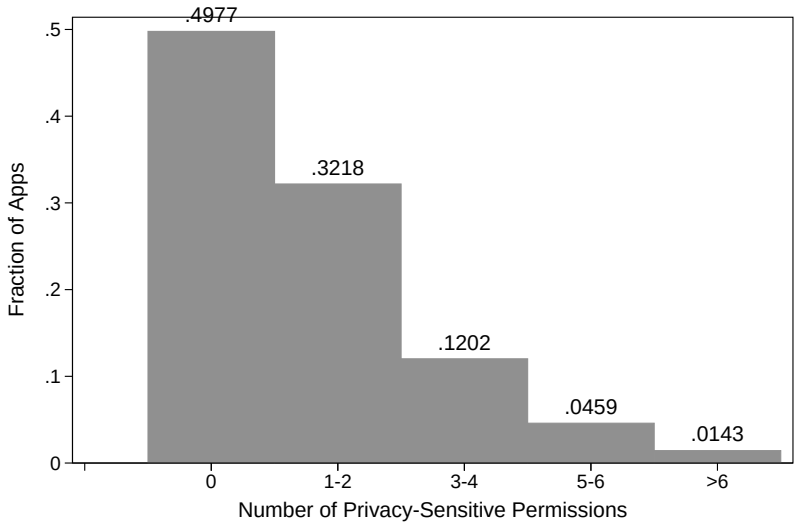
+: Not to mention

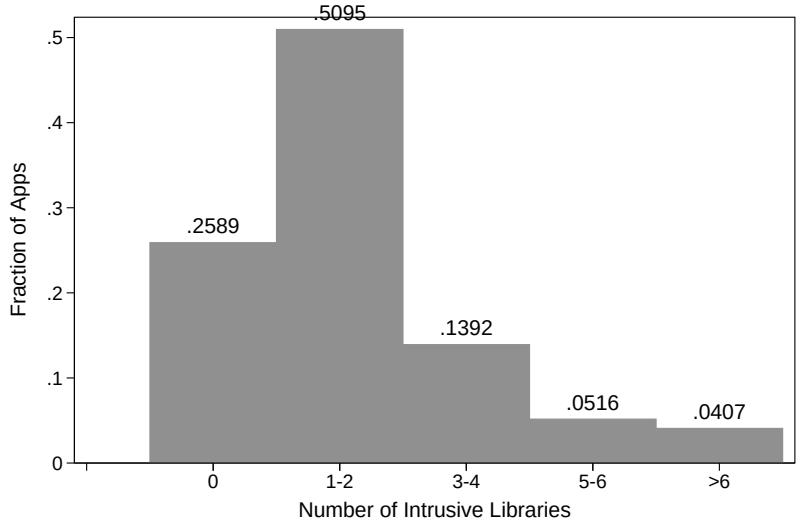
- Rich information on an app's characteristics as covariates

DESCRIPTIVE RESULTS

User Data

Collection and Sharing

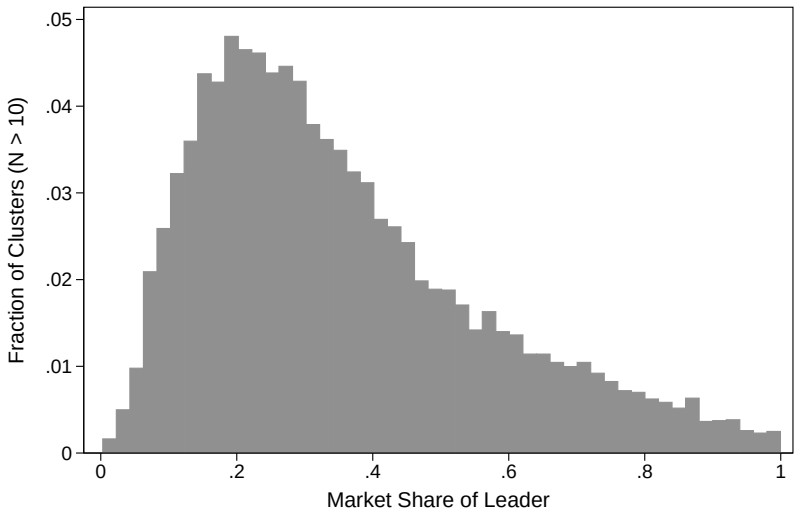


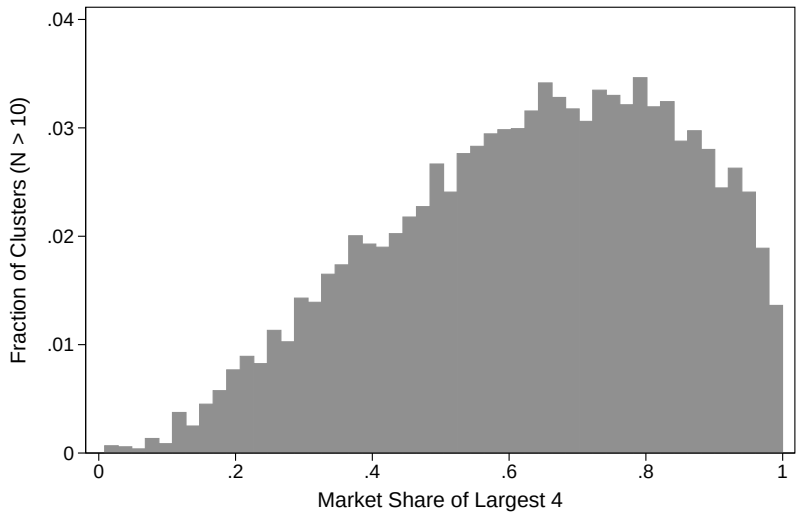


- Half of the apps do not request any privacy-sensitive permission.
- Close to 75 % of the apps share data with a third-party library.

Competition

Strength of Leader and Top 4 in Markets with over 10 Apps

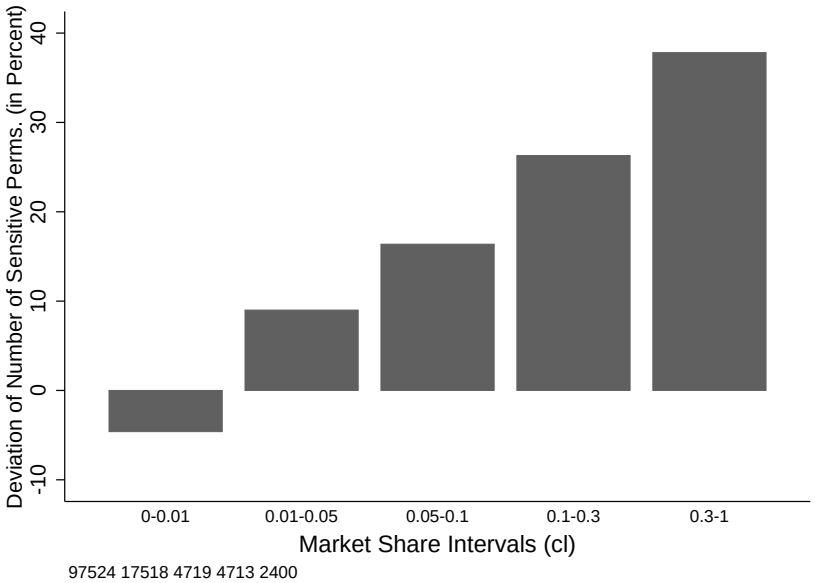


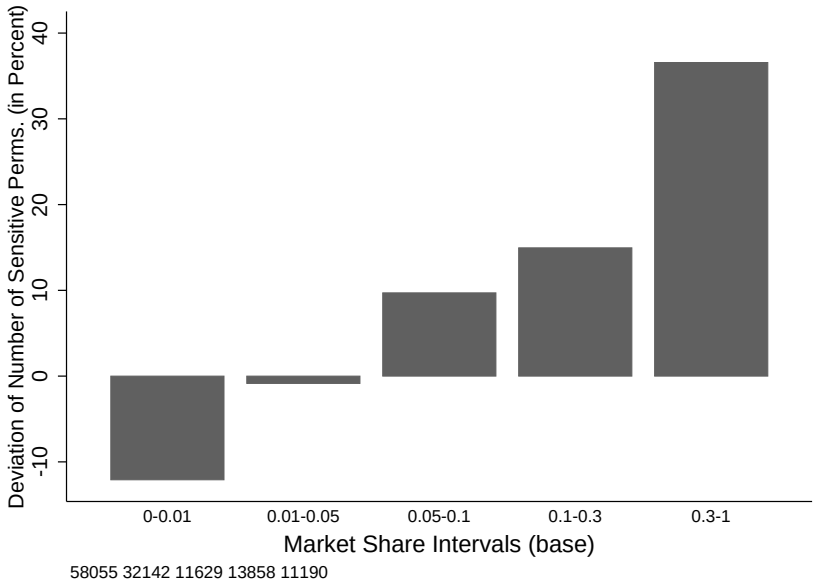


- Majority of Markets has Leaders with a market share over 15%
- Majority of Markets has Top 4 with market share above 50%

Market Share and Data Collection

Descriptive Evidence 2





- On average: apps with a higher market share request more data.

  Note: Contrasting market share intervals and the number of privacy-sensitive permissions.

Summary of our Findings

Results suggest:

1. Substantial share of apps are in highly concentrated markets and data collection/sharing is a common phenomenon.
2. Market concentration/share is positively related to data collection/sharing.
3. Relationship stronger in
    1. important markets and
    2. markets with many players.
4. Leaders collect more data but they don't drive the relationship of interest.

$\rightarrow$ However, effect sizes are very small, once we control for market characteristics etc.

Key Challenges

- One Datawave takes over a week to collect
- Data Cleaning, Data Unification (Website Changes)
- “Holes when an app is missing one quarter”