legislatoR

Political, sociodemographic, and Wikipedia-related data on political elites

Sascha Göbel Simon Munzert

CEAW 2017 | Berlin

December 14th, 2017

Motivation

Why a data package on political elites?

- **continued demand** for individual-level data by researchers, students, analysts, and journalists
- **status quo:** recurrent data collection with the same purpose is **inefficient and restricting**
- **existing data** structures **limited** in scope, hidden **behind paywalls**, or **difficult to access**

Why use & provide Wikipedia data?

- contains archives full of **politicians' biographies**
- widely employed and primary **web information source**
- often deemed **neutral and trustworthy**

Motivation

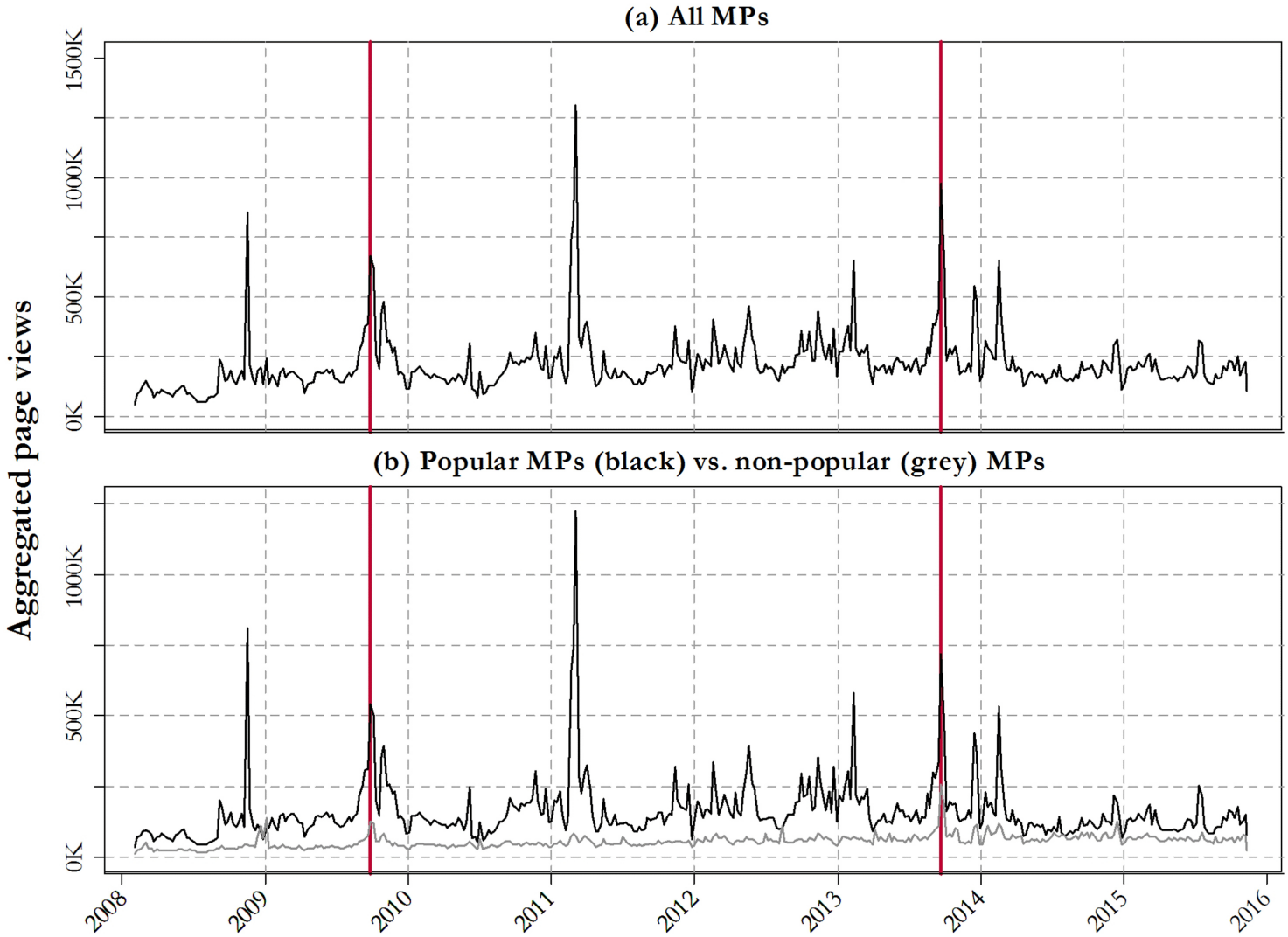
Why a data package on political elites?

- **continued demand** for individual-level data by researchers, students, analysts, and journalists
- **status quo:** recurrent data collection with the same purpose is **inefficient and restricting**
- **existing data** structures **limited** in scope, hidden **behind paywalls**, or **difficult to access**

Why use & provide Wikipedia data?

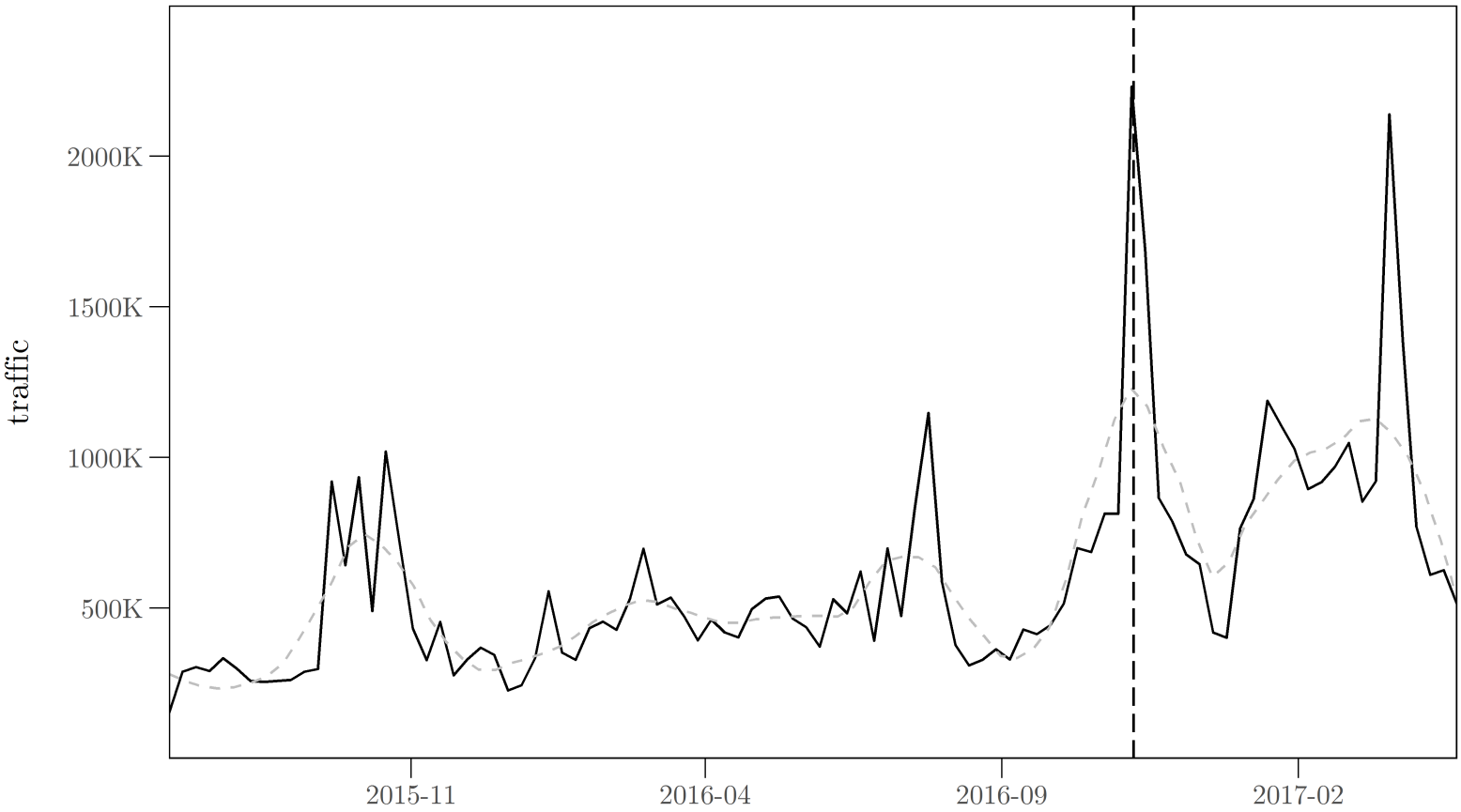
- contains archives full of **politicians' biographies**
- widely employed and primary **web information source**
- often deemed **neutral and trustworthy**

Motivation



Page views of German MPs' Wikipedia entries.

Motivation

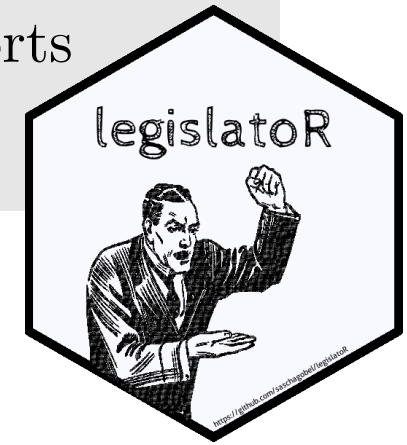


Page views of US Representatives' Wikipedia entries.

Motivation

legislatoR

- resource **efficient**: one stop shop for broad and deep data
- easily **accessible** from R with simple function calls
- facilitates data **integration** and **replication** efforts
- it's **free**



Content and data structure

Content

- **22,917 elected politicians**, all sessions of Austrian Nationalrat, German Bundestag, Irish Dáil, French Assemblée, United States Congress (House and Senate)
- **nine datasets** for each legislature:
  1. *Core* (basic sociodemographic data)
  2. *Political* (basic political data)
  3. *History* (full revision records of Wikipedia biographies)
  4. *Traffic* (daily user traffic on Wikipedia biographies)
  5. *Social* (social media handles and personal website URLs)
  6. *Facial* (URLs to Wikipedia portraits, facial recognition estimates)
  7. *Office* (public offices)
  8. *Occupation* (professions)
  9. *IDs* (identifiers linking to another file, database, or website)

Content and data structure

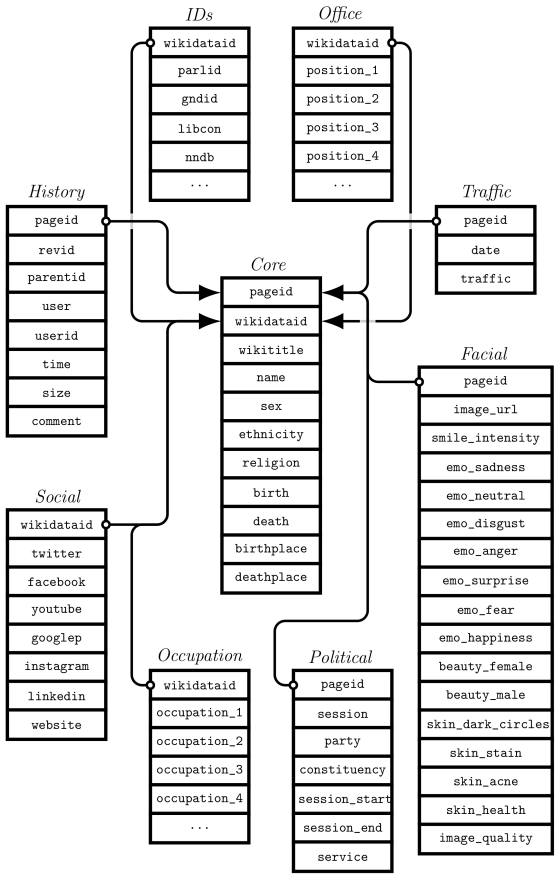
Content

- **22,917 elected politicians**, all sessions of Austrian Nationalrat, German Bundestag, Irish Dáil, French Assemblée, United States Congress (House and Senate)
- **nine datasets** for each legislature:
  1. *Core* (basic sociodemographic data)
  2. *Political* (basic political data)
  3. *History* (full revision records of Wikipedia biographies)
  4. *Traffic* (daily user traffic on Wikipedia biographies)
  5. *Social* (social media handles and personal website URLs)
  6. *Facial* (URLs to Wikipedia portraits, facial recognition estimates)
  7. *Office* (public offices)
  8. *Occupation* (professions)
  9. *IDs* (identifiers linking to another file, database, or website)

Content and data structure

Structure

– **relational** individual-level data – all datasets joinable with *Core* dataset via two keys (Wikipedia page or Wikidata ID)



Content and data structure

Sources

- used **automated** data extraction (XPath, Web APIs)
- Face++ Cognitive Services API, Wikimedia Commons, Wikidata API, Wikipedia, Wikipedia API

Issues

- dependence on data coverage of sources
- in part substantial amount of missings

Content and data structure

Sources

- used **automated** data extraction (XPath, Web APIs)
- Face++ Cognitive Services API, Wikimedia Commons, Wikidata API, Wikipedia, Wikipedia API

Issues

- dependence on data coverage of sources
- in part substantial amount of missings

Usage

Installation

```
> # install from GitHub
> devtools::install_github("saschagobel/legislatoR")
> # load and attach
> library(legislatoR)
```

Get data

- `get_{dataset}()` fetches data from repository
- function takes one argument: `legislature`
- legislature codes: `austria`, `france`, `germany`, `ireland`, `usah`, `usas`

```
> # assign Core dataset for US House of Representatives into environment
> congressmen <- get_core(legislature = "usah")
```

Usage

Installation

```
> # install from GitHub
> devtools::install_github("saschagobel/legislatoR")
> # load and attach
> library(legislatoR)
```

Get data

- `get_{dataset}()` fetches data from repository
- function takes one argument: `legislature`
- legislature codes: `austria`, `france`, `germany`, `ireland`, `usah`, `usas`

```
> # assign Core dataset for US House of Representatives into environment
> congressmen <- get_core(legislature = "usah")
```

Usage

Join and subset data

- data can be joined and subsetted while being fetched
- memory only allocated by parts assigned into environment

```
> # assign Core dataset for US House of Representatives from 115th
> # session into environment
> congressmen115 <- semi_join(x = get_core(legislature = "usah"),
                              y = filter(get_political(legislature =
                                  "usah"), session == 115),
                              by = "pageid")
> # append Wikipedia revision to congressmen115
> congressmen115h <- left_join(x = congressmen115,
                               y = get_history(legislature = "usah"),
                               by = "pageid")
```

Get help

```
> # call legislatoR help file for an overview of function calls
> ?legislatoR
> # call help file for 'History' dataset
> ?get_history
```

Usage

Join and subset data

- data can be joined and subsetted while being fetched
- memory only allocated by parts assigned into environment

```
> # assign Core dataset for US House of Representatives from 115th
> # session into environment
> congressmen115 <- semi_join(x = get_core(legislature = "usah"),
                              y = filter(get_political(legislature =
                                  "usah"), session == 115),
                              by = "pageid")
> # append Wikipedia revision to congressmen115
> congressmen115h <- left_join(x = congressmen115,
                               y = get_history(legislature = "usah"),
                               by = "pageid")
```

Get help

```
> # call legislatoR help file for an overview of function calls
> ?legislatoR
> # call help file for 'History' dataset
> ?get_history
```

Use Cases (1)

- Political advertising on the Wikipedia marketplace of information (Göbel and Munzert Forthcoming)

Existing research

- capitalizes on free accessibility and editability
- political content negotiated (Neff et al. 2013)
- politically biased editing (Kalla and Aronow 2015)

Research question and contribution

- **if**, **how**, and **why** politicians take part in political interaction on Wikipedia?
- study elite behavioral patterns via Wikipedia
- role of the platform for professionalization of campaigning

Use Cases (1)

- Political advertising on the Wikipedia marketplace of information (Göbel and Munzert Forthcoming)

Existing research

- capitalizes on free accessibility and editability
- political content negotiated (Neff et al. 2013)
- politically biased editing (Kalla and Aronow 2015)

Research question and contribution

- **if**, **how**, and **why** politicians take part in political interaction on Wikipedia?
- study elite behavioral patterns via Wikipedia
- role of the platform for professionalization of campaigning

Use Cases (1)

• Wikipedia biographies constitute highly efficient tool for electoral campaigning and political marketing

connect	sway	control
increase online visibility, target voters	advertise, claim credit, take position	shape information, bypass media/party filter, remain undetected

Expectations

Usage and scope differ by incentive to cultivate a personal vote, temporal context, age, party affiliation, popularity, geographical location

Use Cases (1)

• Wikipedia biographies constitute highly efficient tool for electoral campaigning and political marketing

connect	sway	control
increase online visibility, target voters	advertise, claim credit, take position	shape information, bypass media/party filter, remain undetected

Expectations

Usage and scope differ by incentive to cultivate a personal vote, temporal context, age, party affiliation, popularity, geographical location

Use Cases (1)

Empirical strategy

- explore edit histories of Wikipedia biographies
- trace edits originating from parliament's IP range
- explain edit patterns with strategic incentives and sociodemographics

Data

- outcome: count of edits per biography
- predictors: political and sociodemographic MP characteristics and page metrics
- source: scraping of Wikipedia and MPs' Bundestag page HTMLs using XPath and regexp

Use Cases (1)

Empirical strategy

- explore edit histories of Wikipedia biographies
- trace edits originating from parliament's IP range
- explain edit patterns with strategic incentives and sociodemographics

Data

- outcome: count of edits per biography
- predictors: political and sociodemographic MP characteristics and page metrics
- source: scraping of Wikipedia and MPs' Bundestag page HTMLs using XPath and regexp

Use Cases (1)

Case

- German MPs; 2005-2015; 1,100 MPs; 108,775 edits
- third largest Wikipedia, second largest number of edits
- exploit mixed electoral system (district vs. list candidates)

Limitations

- edits cannot be linked to MPs directly
- MPs 'true' editing activity possibly underreported
- actual purpose of edits not revealed by aggregate data

Use Cases (1)

Case

- German MPs; 2005-2015; 1,100 MPs; 108,775 edits
- third largest Wikipedia, second largest number of edits
- exploit mixed electoral system (district vs. list candidates)

Limitations

- edits cannot be linked to MPs directly
- MPs 'true' editing activity possibly underreported
- actual purpose of edits not revealed by aggregate data

Use Cases (1)

Results

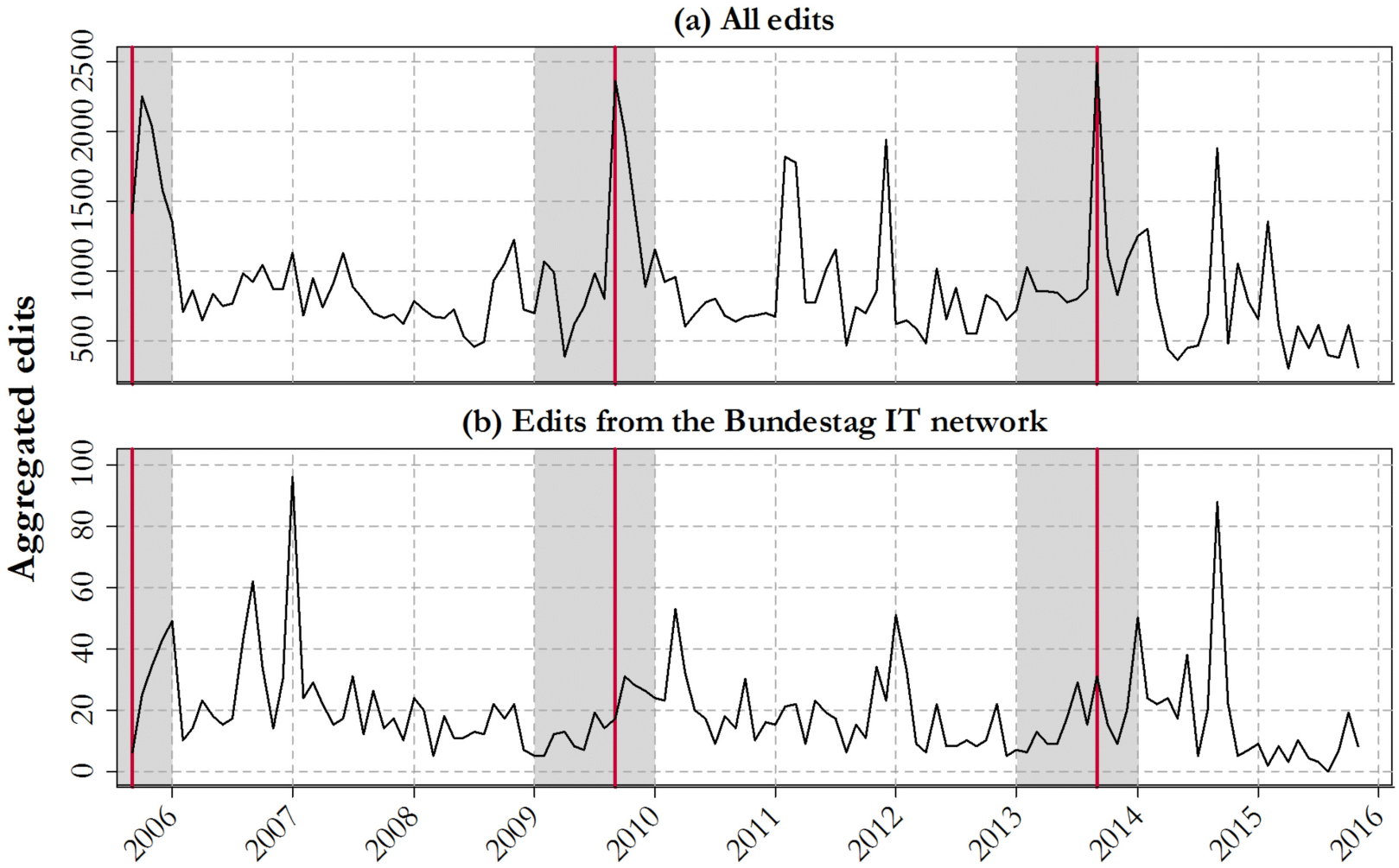
- **substantive public interest** in MPs' Wikipedia entries
- qualitative **evidence** for various kinds **of marketing**
- **51 percent** of the sampled **MP**s exhibit **edits** from the parliament's IP range
- **2.2 percent** ($\approx$ **2400**) **of all edits** can be traced back to the German Bundestag
- **no increase**d activity visible **before elections**
- predictors highlight **strategic incentives**

Use Cases (1)

Typology of edits from parliament's IT network

	category	example
1.	non-substantial edits	correction of tense or typos
2.	private information sharing	information about childhood or housing shared
3.	CV shaping	information about life and career added/removed
4.	provision of campaign information	weblink to campaign page added
5.	highlighting of successes	statements of electoral successes in comparison with competitors
6.	positive reframing	'failures' framed as achievements
7.	removal of adverse information	information about memberships or past life removed

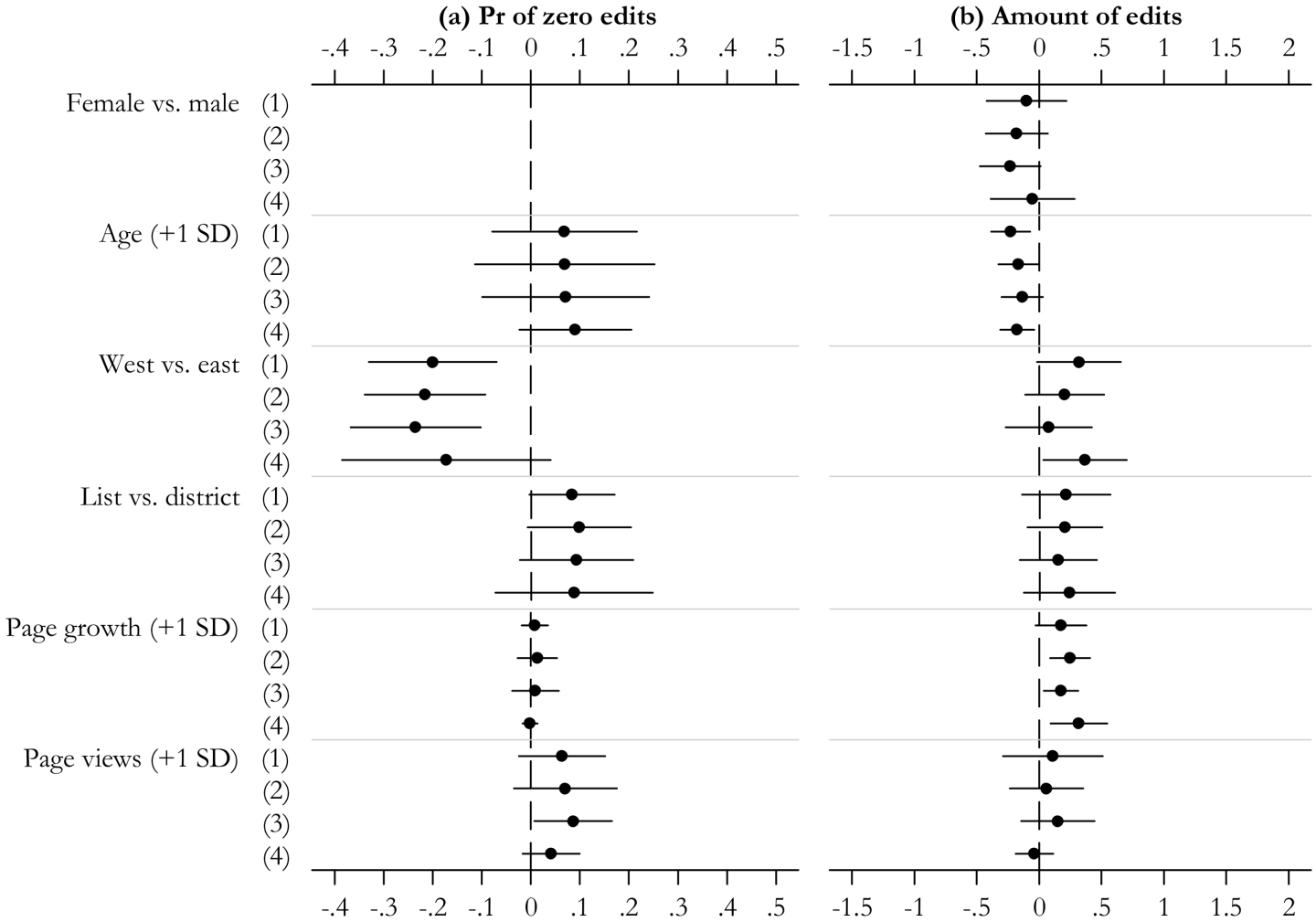
Use Cases (1)



Edits on MPs' Wikipedia biographies over time.

Use Cases (1)

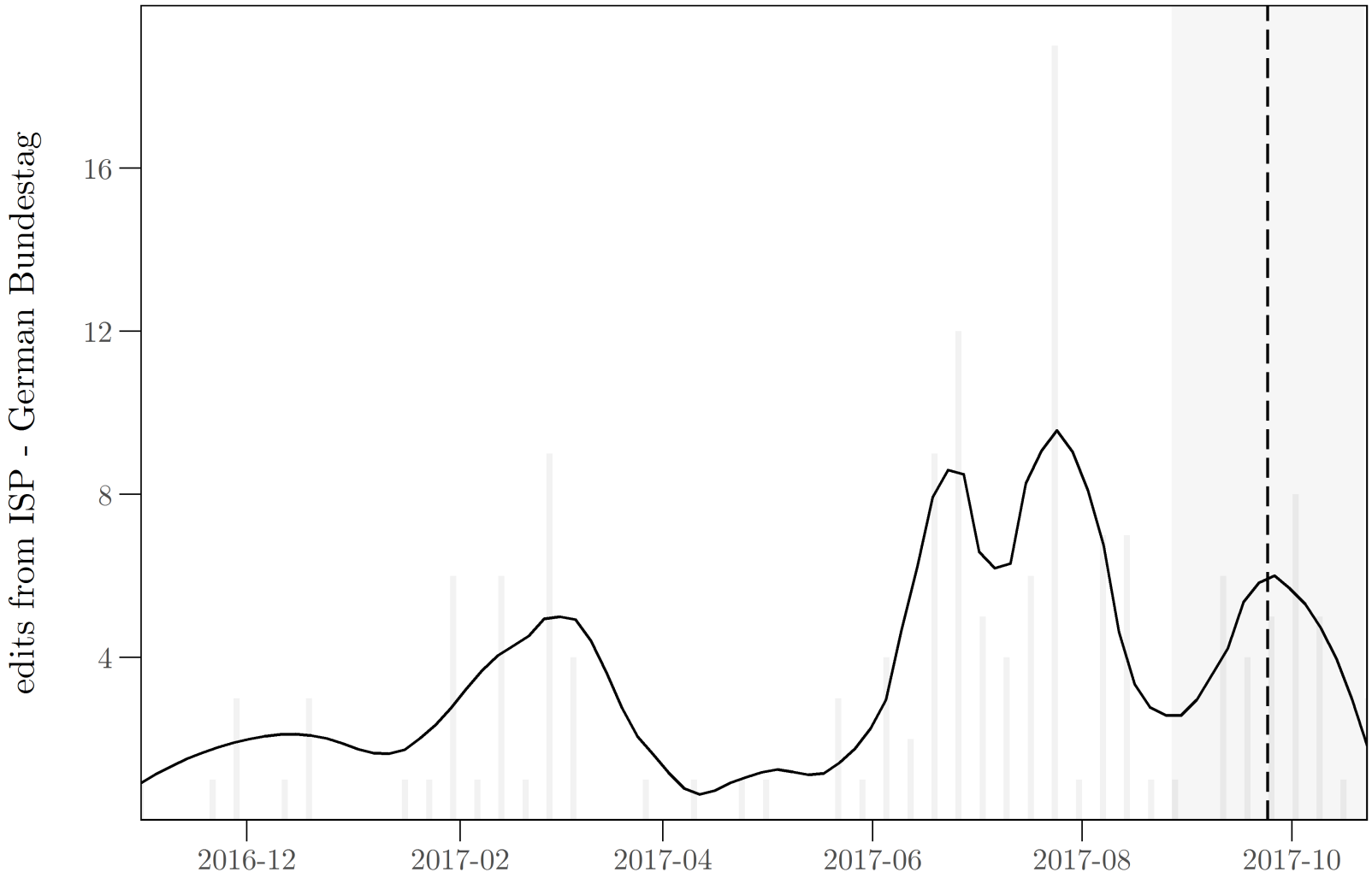
Estimates of factors driving edits from parliament's IP range



Note: Based on ZINB models holding other predictors at observed values.

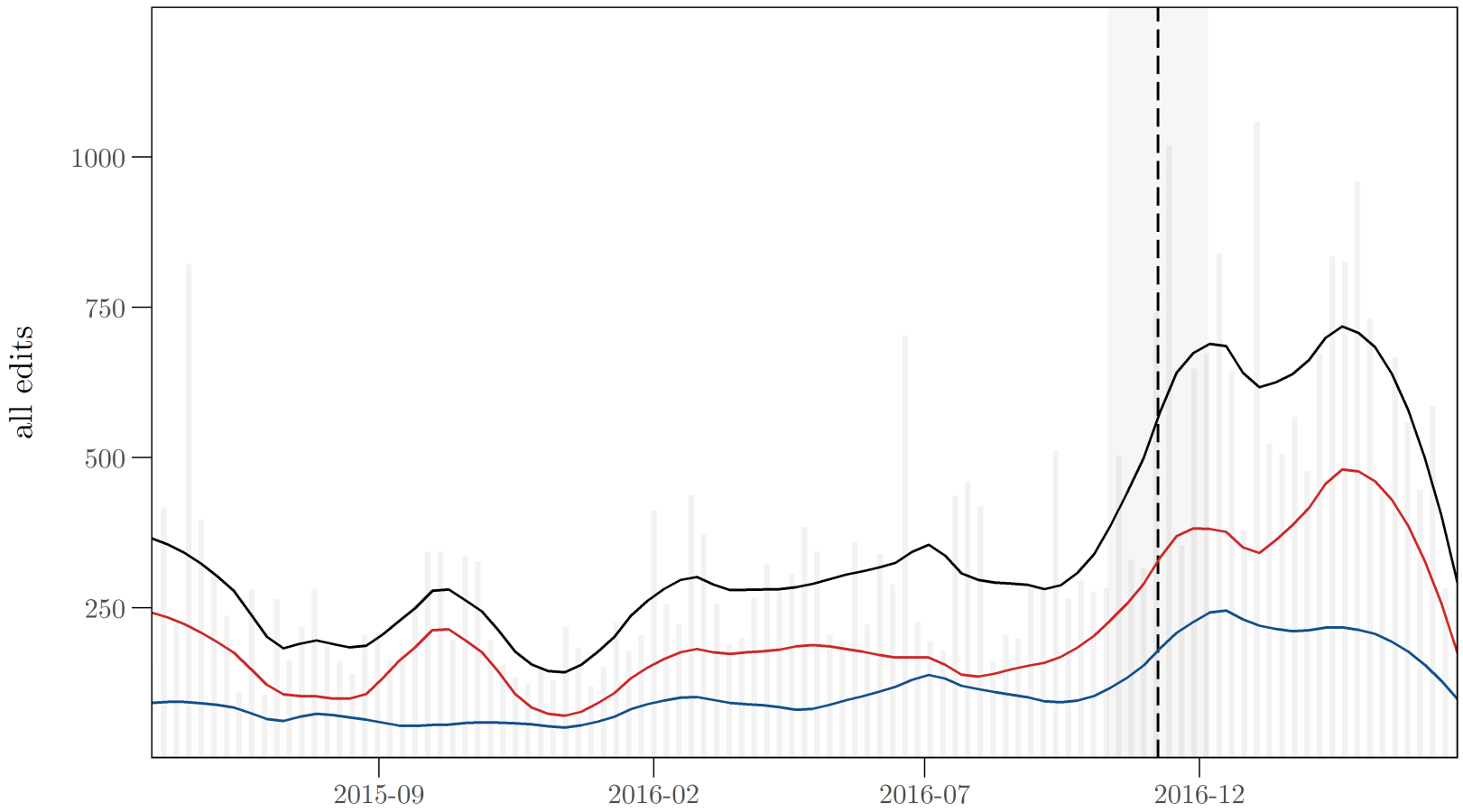
Use Cases (1)

– 18th German Bundestag (continued)

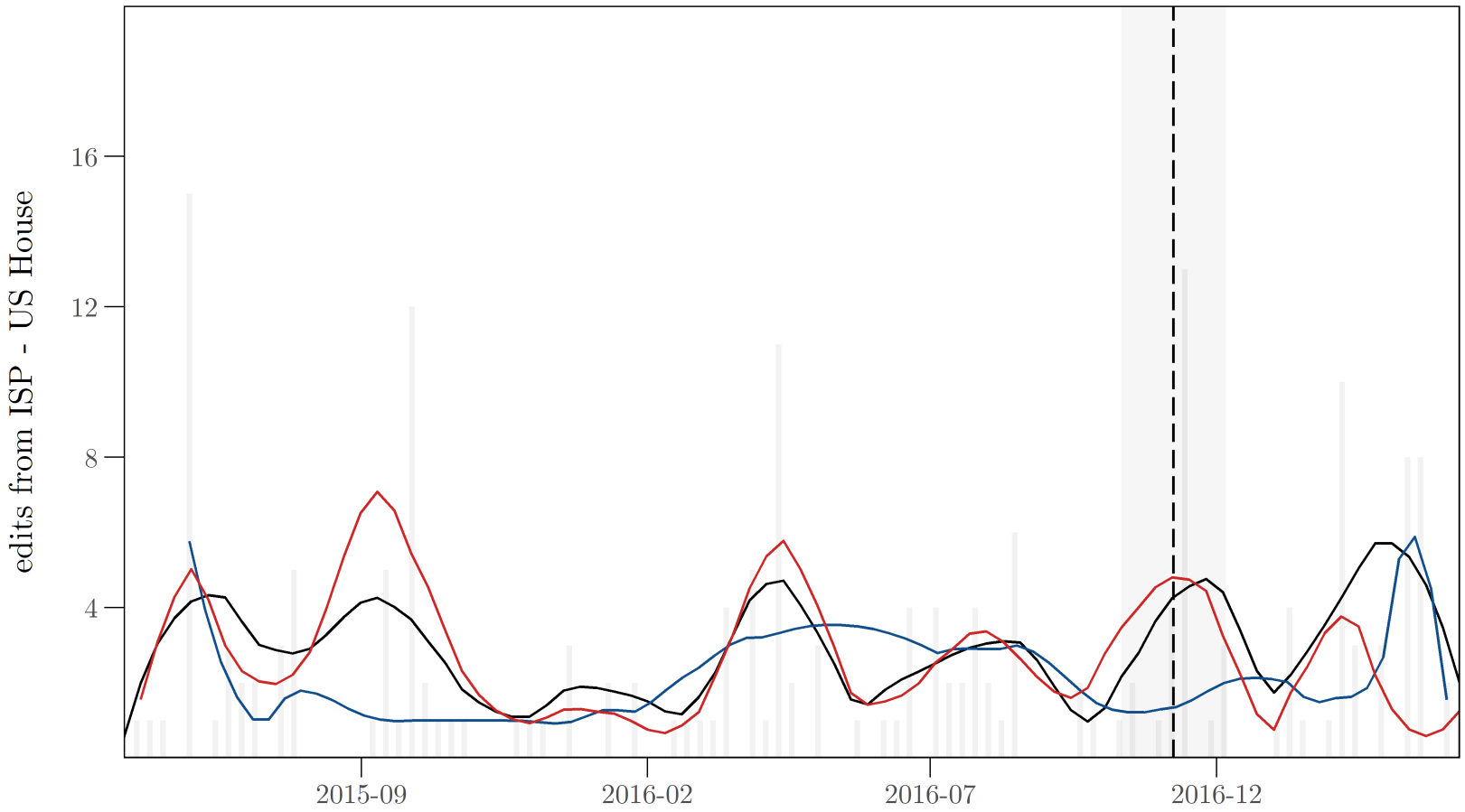


Use Cases (2)

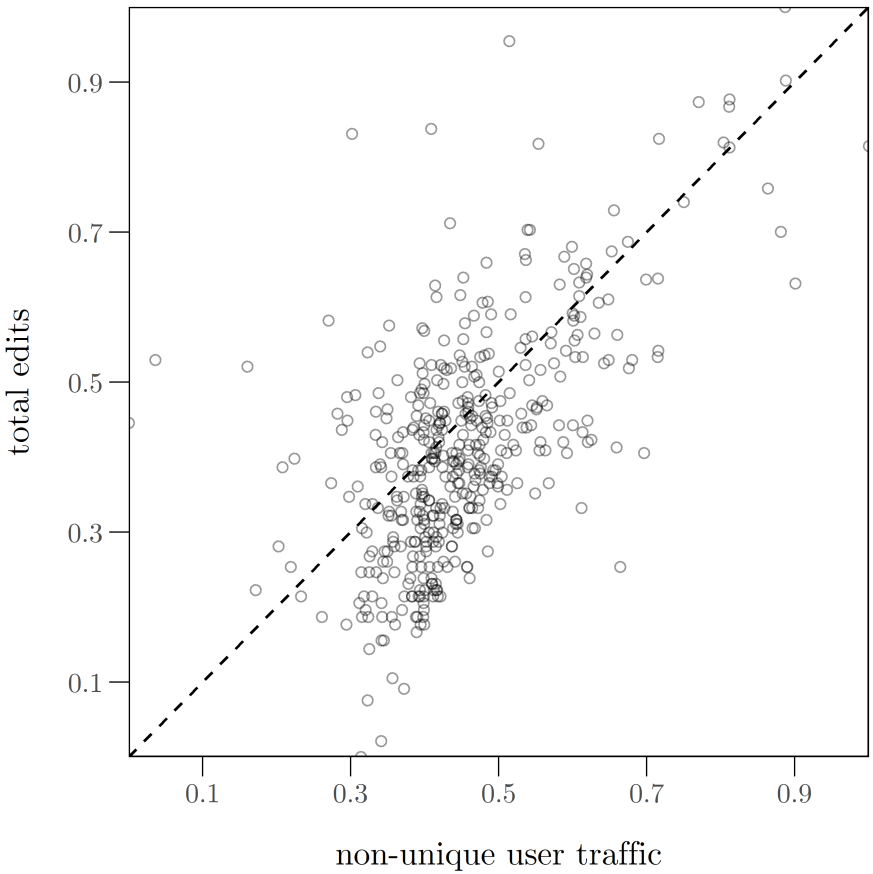
– 114th US House of Representatives

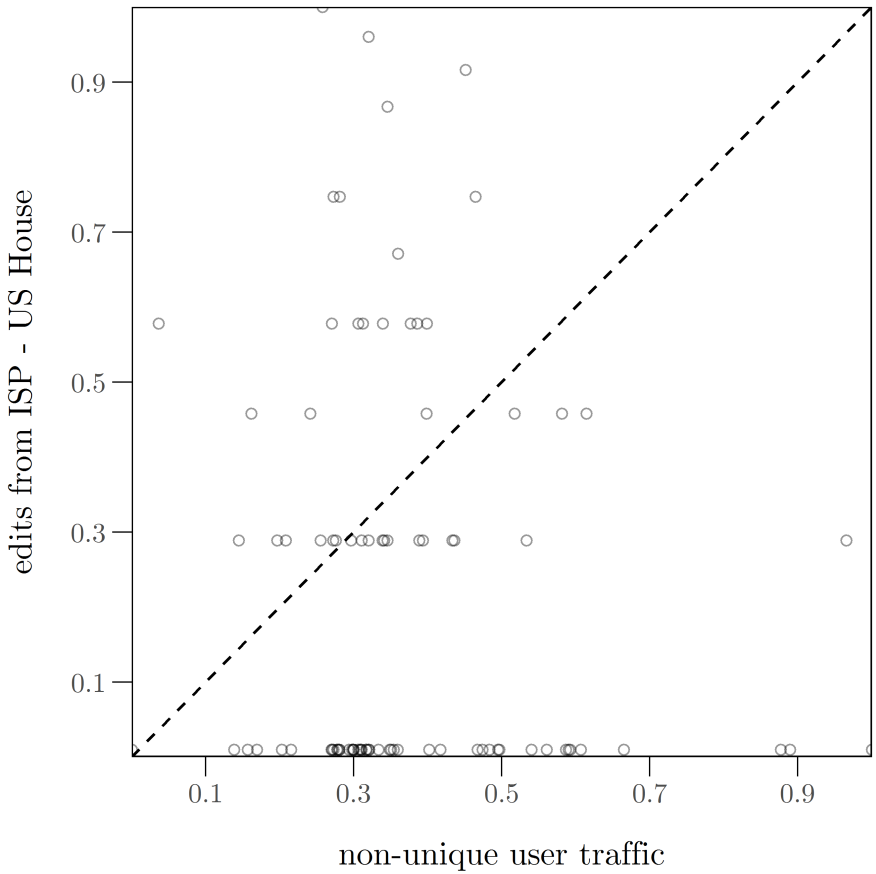


Use Cases (2)

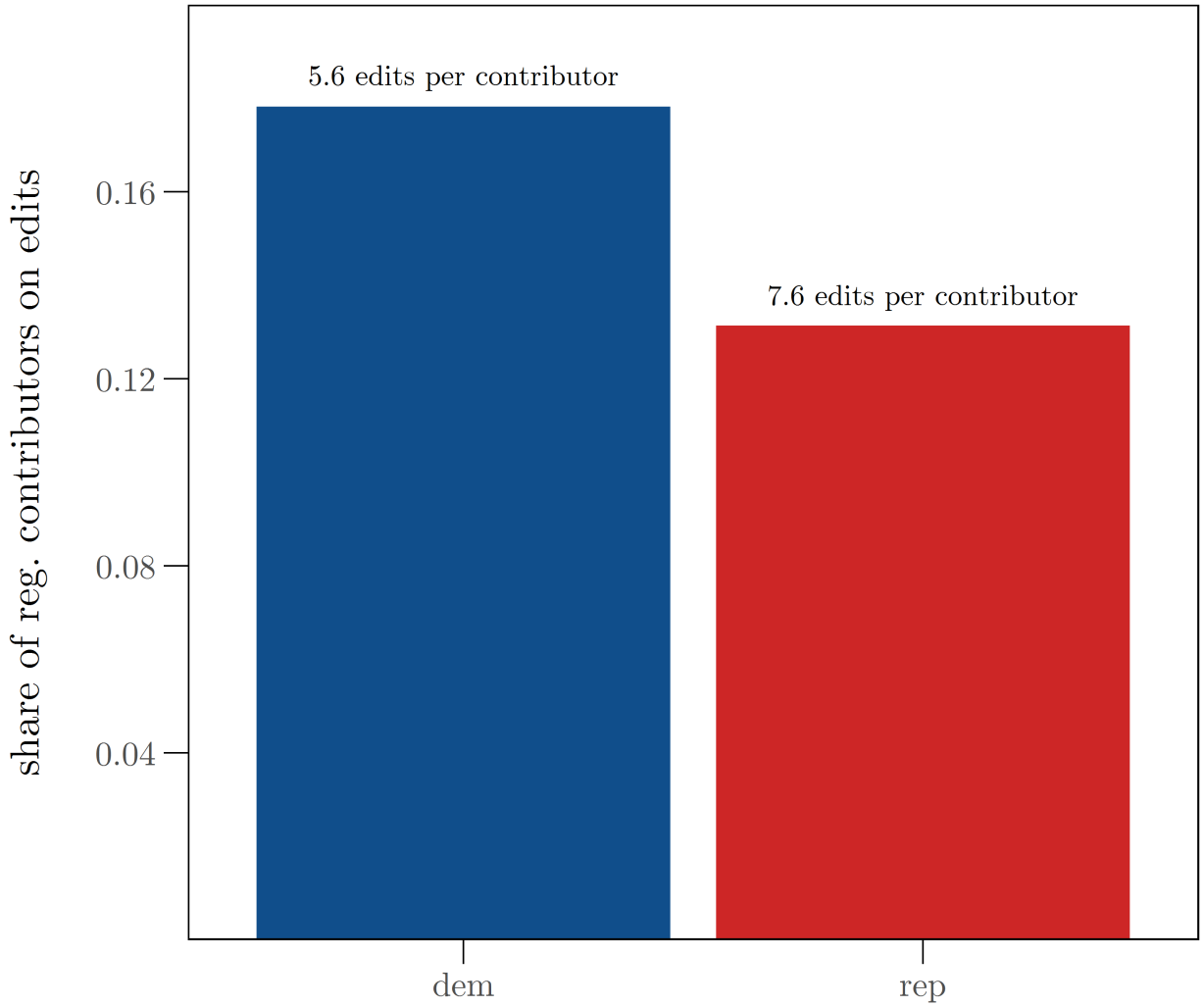


Use Cases (3)

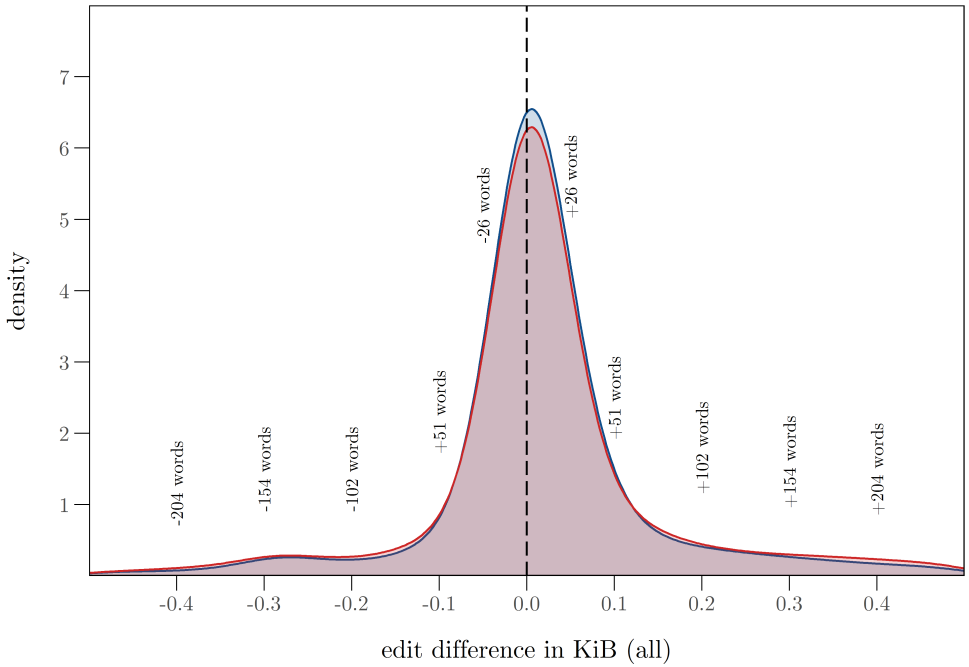

total edits
0.9
0.7
0.5
0.3
0.1
0.1
0.3
0.5
0.7
0.9
non-unique user traffic

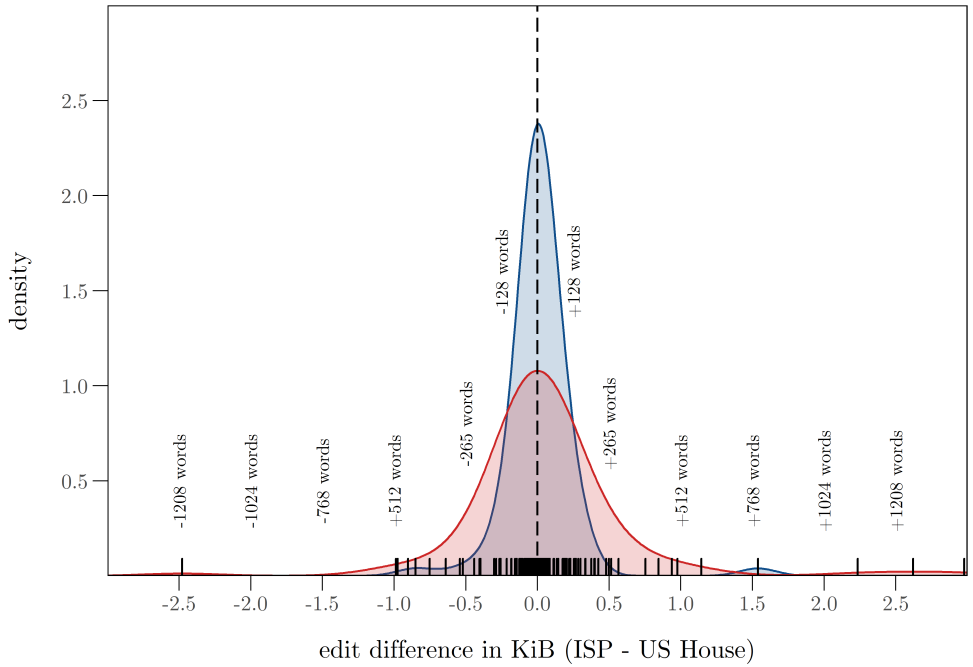

edits from ISP - US House
0.9
0.7
0.5
0.3
0.1
0.1
0.3
0.5
0.7
0.9
non-unique user traffic

Use Cases (4)


5.6 edits per contributor
7.6 edits per contributor
share of reg. contributors on edits
0.16
0.12
0.08
0.04
dem
rep

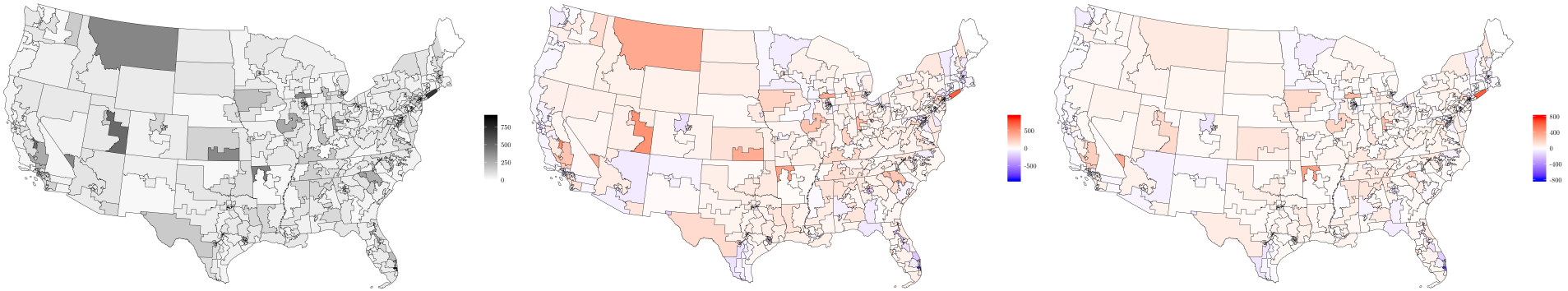
Use Cases (4)


density
-204 words
-154 words
-102 words
+51 words
-26 words
+26 words
+51 words
+102 words
+154 words
+204 words
edit difference in KiB (all)

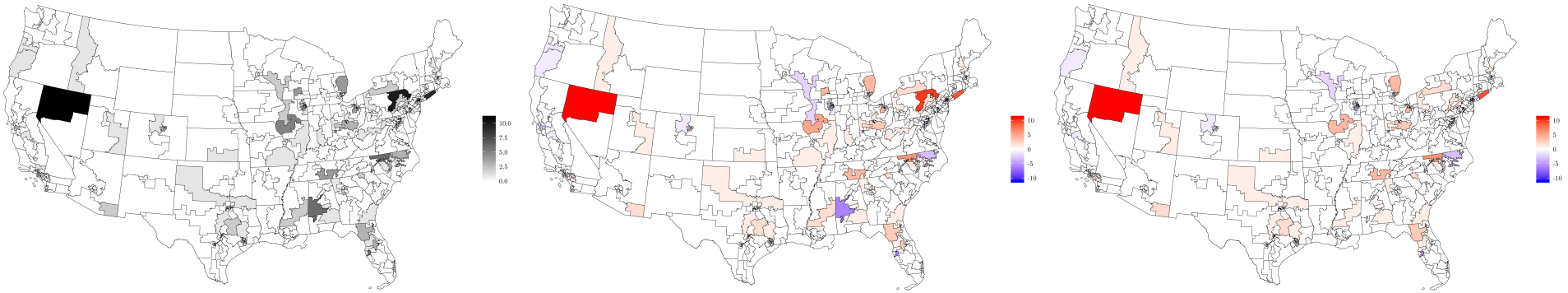

density
-1208 words
-1024 words
-768 words
+512 words
-265 words
-128 words
+128 words
+265 words
+512 words
+768 words
+1024 words
+1208 words
edit difference in KiB (ISP - US House)

Use Cases (5)

all edits

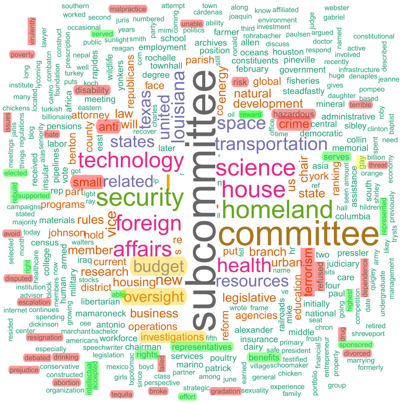


ISP – House

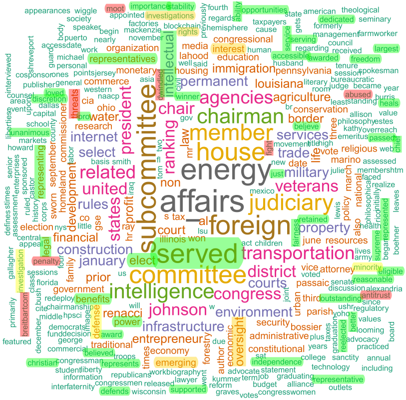


Use Cases (6)

deletions



insertions



-0.4 sentiment +1.3

Thank you for your attention!

Sascha Göbel
sascha.goebel@uni-konstanz.de
Simon Munzert
munzert@hertie-school.org