When survey science met online tracking:
Presenting an error framework for metered data

Oriol J. Bosch | THE LONDON SCHOOL OF ECONOMICS / RECSM-UPF
Melanie Revilla | RECSM-UPF

o.bosch-jover@lse.ac.uk
orioljbosch
https://orioljbosch.com/
When survey science met online tracking:
Presenting an error framework for metered data

Track me but not really:
Device undercoverage and its consequences when tracking online behaviours

Oriol J. Bosch  |  THE LONDON SCHOOL OF ECONOMICS / RECSM-UPF
Melanie Revilla  |  RECSM-UPF

o.bosch-jover@lse.ac.uk
orioljbosch
https://orioljbosch.com/
INTRODUCTION

Tracking online behaviours using a meter

Definition

*Metered data* is obtained from a meter willingly installed or configured by a sample of participants on their devices (PCs, tablets and/or smartphones).

A *meter* refers to a heterogeneous group of tracking technologies that allow sharing with the researchers, at least, *information about the URLs of the web pages visited by the participants*. 
**Definition**

*Metered data* is obtained from a meter willingly installed or configured by a sample of participants on their devices (PCs, tablets and/or smartphones).

A *meter* refers to a heterogeneous group of tracking technologies that allow sharing with the researchers, at least, *information about the URLs of the web pages visited by the participants.*
INTRODUCTION

Tracking online behaviours using a meter

Benefits of metered data

• Objective and free of recall errors

• Continuously collected in real time

• Pre-designed sample of participants
Metered data in past research

33 papers identified, 26 since 2019
BACKGROUND

Metered data in past research

33 papers identified, 26 since 2019
Predicting Voting Behavior Using Digital Trace Data

Ruben L. Bach¹, Christoph Kern¹, Ashley Amaya², Florian Keusch¹, Frauke Kreuter¹,³,⁴, Jan Hecht³, and Jonathan Heineman⁵

Abstract
A major concern arising from ubiquitous personal online activity is that substantial numbers of individuals’ online activity can immediately but short-lived increases in website visits and knowledge of recent events. After adjusting for multiple comparisons, however, we find little evidence of a direct impact on opinions or affect. Still, results from larger survey waves suggest that both treatments produce a lasting and meaningful decrease in trust in the mainstream media up to 1 year later. Consistent with the minimal-effects tradition, direct consequences of online partisan media are limited, although our findings raise questions about the possibility of subtle, cumulative dynamics. The combination of experimental and computational social science techniques illustrates a powerful approach for studying the long-term consequences of exposure to partisan news.

The consequences of online partisan media

Andrew M. Guess¹,²,³,⁴, Pablo Barbera¹,², Simon Munzer⁵,⁶, and Jung-Hwan Yang⁷,⁸,⁹,¹⁰

¹Department of Politics, Princeton University, Princeton, NJ 08544; ²School of Public and International Affairs, Princeton University, Princeton, NJ 08544; ³Department of Political Science and International Relations, University of Southern California, Los Angeles, CA 90089; ⁴Data Science Lab, Hertie School, 10117 Berlin, Germany; and ⁵Department of Communication, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Is Facebook Eroding the Public Agenda? Evidence From Survey and Web-Tracking Data

Ana S. Cardenal¹, Carol Galais², and Silvia Majó-Vázquez³

¹School of Law and Political Science, Universitat Oberta de Catalunya, Spain; ²Political Science and Public Law Department, Universitat Autònoma de Barcelona, Spain; ³Reuter Institute for the Study of Journalism, University of Oxford, UK

BACKGROUND

BACKGROUND

The volume of individuals’ online activity is that changing may

ARTICLE INFO

Article info
Article history:
Received 27 June 2020
Revised 1 October 2020
Accepted 7 October 2020
Available online 22 October 2020

Keywords:
Vaccine hesitancy
Vaccine deprivation
Online search
Social media

Abstract
What role do ideologically extreme media play in the polariz-
ation of society? Here we report results from a randomized longitudinal field experiment embedded in a nationally representative online panel survey (N = 1,037) in which participants were incentivized to change their browser default settings and social media following patterns, boosting the likelihood of encountering news with either a left-leaning ( HuffPost ) or right-leaning ( Fox News ) slant during the 2018 US midterm election campaign. Data on ≈ 19 million web visits by respondents indicate that resulting changes in news consumption persisted for at least 8 wk. Greater exposure to partisan news can cause immediate but short-lived increases in website visits and knowledge of recent events. After adjusting for multiple comparisons, however, we find little evidence of a direct impact on opinions or affect. Still, results from larger survey waves suggest that both treatments produce a lasting and meaningful decrease in trust in the mainstream media up to 1 year later. Consistent with the minimal-effects tradition, direct consequences of online partisan media are limited, although our findings raise questions about the possibility of subtle, cumulative dynamics. The combination of experimental and computational social science techniques illustrates a powerful approach for studying the long-term consequences of exposure to partisan news.

ARGUMENTS

Is Facebook Eroding the Public Agenda? Evidence From Survey and Web-Tracking Data

Ana S. Cardenal¹, Carol Galais², and Silvia Majó-Vázquez³

¹School of Law and Political Science, Universitat Oberta de Catalunya, Spain; ²Political Science and Public Law Department, Universitat Autònoma de Barcelona, Spain; ³Reuter Institute for the Study of Journalism, University of Oxford, UK

BACKGROUND

BACKGROUND

The volume of individuals’ online activity is that changing may

ARTICLE INFO

Article info
Article history:
Received 27 June 2020
Revised 1 October 2020
Accepted 7 October 2020
Available online 22 October 2020

Keywords:
Vaccine hesitancy
Vaccine deprivation
Online search
Social media

Abstract
What role do ideologically extreme media play in the polariz-
ation of society? Here we report results from a randomized longitudinal field experiment embedded in a nationally representative online panel survey (N = 1,037) in which participants were incentivized to change their browser default settings and social media following patterns, boosting the likelihood of encountering news with either a left-leaning ( HuffPost ) or right-leaning ( Fox News ) slant during the 2018 US midterm election campaign. Data on ≈ 19 million web visits by respondents indicate that resulting changes in news consumption persisted for at least 8 wk. Greater exposure to partisan news can cause immediate but short-lived increases in website visits and knowledge of recent events. After adjusting for multiple comparisons, however, we find little evidence of a direct impact on opinions or affect. Still, results from larger survey waves suggest that both treatments produce a lasting and meaningful decrease in trust in the mainstream media up to 1 year later. Consistent with the minimal-effects tradition, direct consequences of online partisan media are limited, although our findings raise questions about the possibility of subtle, cumulative dynamics. The combination of experimental and computational social science techniques illustrates a powerful approach for studying the long-term consequences of exposure to partisan news.

ARGUMENTS

Is Facebook Eroding the Public Agenda? Evidence From Survey and Web-Tracking Data

Ana S. Cardenal¹, Carol Galais², and Silvia Majó-Vázquez³

¹School of Law and Political Science, Universitat Oberta de Catalunya, Spain; ²Political Science and Public Law Department, Universitat Autònoma de Barcelona, Spain; ³Reuter Institute for the Study of Journalism, University of Oxford, UK

BACKGROUND

BACKGROUND

The volume of individuals’ online activity is that changing may

ARTICLE INFO

Article info
Article history:
Received 27 June 2020
Revised 1 October 2020
Accepted 7 October 2020
Available online 22 October 2020

Keywords:
Vaccine hesitancy
Vaccine deprivation
Online search
Social media

Abstract
What role do ideologically extreme media play in the polariz-
ation of society? Here we report results from a randomized longitudinal field experiment embedded in a nationally representative online panel survey (N = 1,037) in which participants were incentivized to change their browser default settings and social media following patterns, boosting the likelihood of encountering news with either a left-leaning ( HuffPost ) or right-leaning ( Fox News ) slant during the 2018 US midterm election campaign. Data on ≈ 19 million web visits by respondents indicate that resulting changes in news consumption persisted for at least 8 wk. Greater exposure to partisan news can cause immediate but short-lived increases in website visits and knowledge of recent events. After adjusting for multiple comparisons, however, we find little evidence of a direct impact on opinions or affect. Still, results from larger survey waves suggest that both treatments produce a lasting and meaningful decrease in trust in the mainstream media up to 1 year later. Consistent with the minimal-effects tradition, direct consequences of online partisan media are limited, although our findings raise questions about the possibility of subtle, cumulative dynamics. The combination of experimental and computational social science techniques illustrates a powerful approach for studying the long-term consequences of exposure to partisan news.
How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use
Theo Araujo, Anke Wonneberger, Peter Neijens, and Claes de Vreese
Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT
Given the importance of survey measures of online media use for communication research, it is crucial to assess and improve their quality, in particular because the increasingly fragmented and ubicompics of the internet complicates the accuracy of surveys. This article contributes to the discussion regarding survey design by presenting relevant factors and by testing survey design strategies. Data from tracking data and survey data are compared. The accuracy of survey-based estimates of Internet usage is confirmed low levels of accuracy. The article also explores biases due to a range of factors, including (actual) Internet usage, proper usage of mobile devices, and an increase in the reporting of self-reported Internet use in survey data.

Two Half-Truths Make a Whole? On Bias in Self-Reports and Tracking Data
Pascal Jürgens¹, Birgit Stark¹, and Melanie Magin²

ABSTRACT
The vast majority of empirical research on online communication, or media use in general, relies on self-report measures instead of behavioral data. Previous research has shown that the accuracy of these self-report measures is quite low, and both over- and underreporting of media use are commonplace. This study compares self-reports of Internet use with client logs from a large household sample. Results show that the accuracy of self-reported frequency and duration of Internet use is quite low, and that survey data are only moderately correlated with log file data. Moreover, there are systematic patterns of misreporting, especially overreporting, rather than random deviations from the log files. Self-reports for specific content such as social network sites or video platforms seem to be more accurate and less consistently biased than self-reports of generic frequency or duration of Internet use. The article closes by demonstrating the consequences of biased self-reports and discussing possible solutions to the problem.
Inferences for finite populations

**Metered data can potentially suffer from different types of errors**

Shared devices and observation of only part of the activity

- 60% of desktops, 40% of laptops and tablets, and 9% of smartphones shared to some degree (Revilla et al., 2017)
- 28% with the meter installed in all devices (Pew Research Center, 2020)

Technical issues and reactivity / social desirability bias (Jurgens et al., 2020; Toth and Trifonova, 2020)

Substantive conclusions vary depending on what is considered as a visit (3 seconds / 30 seconds / 120 seconds) (Mangold et al., 2021)
Inferences for finite populations

Metered data can potentially suffer from different types of errors

Shared devices and observation of only part of the activity

• 60% of desktops, 40% of laptops and tablets, and 9% of smartphones shared to some degree (Revilla et al., 2017)

• 28% with the meter installed in all devices (Pew Research Center, 2020)

Technical issues and reactivity / social desirability bias (Jurgens et al., 2020; Toth and Trifonova, 2020)

Substantive conclusions vary depending on what is considered as a visit (3 seconds / 30 seconds / 120 seconds) (Mangold et al., 2021)
Inferences for finite populations

Metered data can potentially suffer from different types of errors

Shared devices and observation of only part of the activity

- 60% of desktops, 40% of laptops and tablets, and 9% of smartphones shared to some degree (Revilla et al., 2017)
- 28% with the meter installed in all devices (Pew Research Center, 2020)

Technical issues and reactivity / social desirability bias (Jurgens et al., 2020; Toth and Trifonova, 2020)

Substantive conclusions vary depending on what is considered as a visit (3 seconds / 30 seconds / 120 seconds) (Mangold et al., 2021)
BACKGROUND

Inferences for finite populations

Metered data can potentially suffer from different types of errors

Shared devices and observation of only part of the activity

- 60% of desktops, 40% of laptops and tablets, and 9% of smartphones shared to some degree (Revilla et al., 2017)
- 28% with the meter installed in all devices (Pew Research Center, 2020)

Technical issues and reactivity / social desirability bias (Jurgens et al., 2020; Toth and Trifonova, 2020)

Substantive conclusions vary depending on what is considered as a visit (3 seconds / 30 seconds / 120 seconds) (Mangold et al., 2021)

A systematic **categorization** and **conceptualization** of metered data errors

Nor empirical demonstrations of (many of) those errors!
Main goals and contribution

Total Error Framework for metered data

• #1 **Summarize** the data collection and analysis process for metered data.

• #2 **Conceptualize and categorize** all errors components (e.g. measurement errors) and causes (e.g. social desirability) that can occur when using metered data.
Main goals and contribution

Total Error Framework for metered data

• #1 **Summarize** the data collection and analysis process for metered data.

• #2 **Conceptualize and categorize** all errors components (e.g. measurement errors) and causes (e.g. social desirability) that can occur when using metered data.

1) Choose the best design options for metered data.
2) Make better informed decisions while planning when and how to supplement or replace survey data with metered data.
3) Help assess research using metered data.
Main goals and contribution

Total Error Framework for metered data

• #1 **Summarize** the data collection and analysis process for metered data.

• #2 **Conceptualize and categorize** all errors components (e.g. measurement errors) and causes (e.g. social desirability) that can occur when using metered data.

Adapting instead of reinventing

• Follow approach by Amaya et al (2020) with their Total Error Framework for Big Data

• 7 error components of the TSE (Groves et al., 2009) as starting point:
  • Coverage errors, sampling errors, missing data errors, adjustment errors, specification errors, measurement errors and processing errors
RESULTS

Data collection and analysis process
RESULTS

Data collection and analysis process

Define concept of interest

Average hours of consumption of online political news
RESULTS

Data collection and analysis process

- Define concept of interest
- Design measurement
- Average hours of consumption of online political news
- Average time recorded of the visits to online political outlets’ URLs.
Data collection and analysis process

Define concept of interest

Design measurement

Develop/choose the technology

Average hours of consumption of online political news

Average time recorded of the visits to online political outlets’ URLs.

Proxy for IOS/App for others
RESULTS

Data collection and analysis process

- Define concept of interest
- Design measurement
- Develop/choose the technology

Average hours of consumption of online political news
Average time recorded of the visits to online political outlets’ URLs.
Proxy for iOS/ App for others

Define target inferential population
People living in the UK older than 18
RESULTS

Data collection and analysis process

- Define concept of interest
- Define target inferential population
- Design measurement
- Construct frame
- Develop/choose the technology
- Proxy for iOS/App for others

Average hours of consumption of online political news
Average time recorded of the visits to online political outlets' URLs.

People living in the UK older than 18
Postal Address Frame
RESULTS

Data collection and analysis process

1. Define concept of interest
2. Define target inferential population
3. Design measurement
4. Construct frame
5. Develop/choose the technology
6. Draw sample

- Simple Random Sampling
- People living in the UK older than 18
- Postal Address Frame
- Average hours of consumption of online political news
- Average time recorded of the visits to online political outlets' URLs.
- Proxy for IOS/ App for others
RESULTS

Data collection and analysis process

- Define concept of interest
- Design measurement
- Develop/choose the technology
- Draw sample
- Install the meter

Simple Random Sampling

Proxy for IOS/ App for others

Average hours of consumption of online political news

Average time recorded of the visits to online political outlets' URLs.

Define target inferential population

People living in the UK older than 18

Postal Address Frame
RESULTS

Data collection and analysis process

- Define concept of interest
- Design measurement
- Develop/choose the technology
- Draw sample
- Install the meter
- Identify/generate data source

Population:
- Define target inferential population: People living in the UK older than 18
- Postal Address Frame

Data:
- Average hours of consumption of online political news
- Average time recorded of the visits to online political outlets' URLs
- Proxy for IOS/ App for others

Methods:
- Simple Random Sampling
Data collection and analysis process

- Define concept of interest
- Define target inferential population
- Design measurement
- Construct frame
- Develop/choose the technology
- Draw sample
- Install the meter
- Identify/generate data source
  - Extract
  - Transform
  - Load
- Simple Random Sampling
- Only information of news sites from UK
  - Code ideology of the content of articles
  - Load and store on server or device

- Average hours of consumption of online political news
- Average time recorded of the visits to online political outlets' URLs.
- Proxy for IOS/App for others
- People living in the UK older than 18
  - Postal Address Frame
RESULTS

Data collection and analysis process

Define concept of interest
Define target inferential population
Design measurement
Construct frame
Develop / choose the technology
Draw sample
Install the meter
Identify / generate data source
Extract
Transform
Load
Model

Average hours of consumption of online political news
Average time recorded of the visits to online political outlets' URLs.
Proxy for IOS/ App for others

Only information of news sites from UK
Code ideology of the content of articles
Load and store on server or device
Weight / Imputation

People living in the UK older than 18
Postal Address Frame
Simple Random Sampling

RESULTS
RESULTS

Data collection and analysis process

Define concept of interest
Define target inferential population
Design measurement
Construct frame
Develop/choose the technology
Draw sample
Install the meter
Identify/generate data source
Extract
Transform
Load
Model
Create estimates

- Average hours of consumption of online political news
- Average time recorded of the visits to online political outlets’ URLs.
- Proxy for IOS/App for others

- People living in the UK older than 18
- Postal Address Frame
- Simple Random Sampling

- Only information of news sites from UK
- Code ideology of the content of articles
- Load and store on server or device
- Weight/Imputation

- Only information of news sites from UK
- Code ideology of the content of articles
- Load and store on server or device
- Weight/Imputation
RESULTS

Data collection and analysis process

- Define concept of interest
- Define target inferential population
- Design measurement
- Construct frame
- Develop/choose the technology
- Draw sample
- Install the meter
- Identify/generate data source
- Extract
- Transform
- Load
- Model
- Create estimates

Errors:
- Specification error
- Measurement error
- Sampling error
- Missing data error
- Adjustment error
- Coverage error
- Processing error
- Measurement error
## Error components and their causes

<table>
<thead>
<tr>
<th>Error components</th>
<th>Specific error causes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specification error</td>
<td>- Measuring concepts from which not enough data is available</td>
</tr>
<tr>
<td></td>
<td>- Inferring attitudes</td>
</tr>
<tr>
<td></td>
<td>- Defining valid information</td>
</tr>
<tr>
<td>Measurement error</td>
<td>- Non-trackable target</td>
</tr>
<tr>
<td></td>
<td>- Meter not installed</td>
</tr>
<tr>
<td></td>
<td>- Uninstalling the meter</td>
</tr>
<tr>
<td></td>
<td>- New non-tracked device</td>
</tr>
<tr>
<td></td>
<td>- Technology limitations</td>
</tr>
<tr>
<td></td>
<td>- Technology errors</td>
</tr>
<tr>
<td></td>
<td>- Hidden behaviours</td>
</tr>
<tr>
<td></td>
<td>- Shared device</td>
</tr>
<tr>
<td></td>
<td>- Social desirability</td>
</tr>
<tr>
<td></td>
<td>- Extraction error</td>
</tr>
<tr>
<td>Processing error</td>
<td>- Coding error</td>
</tr>
<tr>
<td></td>
<td>- Aggregation at the domain level</td>
</tr>
<tr>
<td></td>
<td>- Data anonymization</td>
</tr>
<tr>
<td>Coverage error</td>
<td>- Non-trackable individuals</td>
</tr>
<tr>
<td>Sampling error</td>
<td>- Same error causes than for surveys</td>
</tr>
<tr>
<td>Missing data error</td>
<td>- Noncontact</td>
</tr>
<tr>
<td></td>
<td>- Non-consent</td>
</tr>
<tr>
<td></td>
<td>- Non-trackable target</td>
</tr>
<tr>
<td></td>
<td>- Meter not installed</td>
</tr>
<tr>
<td></td>
<td>- Uninstalling the meter</td>
</tr>
<tr>
<td></td>
<td>- New non-tracked device</td>
</tr>
<tr>
<td></td>
<td>- Technology limitations</td>
</tr>
<tr>
<td></td>
<td>- Technology error</td>
</tr>
<tr>
<td></td>
<td>- Hidden behaviour</td>
</tr>
<tr>
<td></td>
<td>- Social desirability</td>
</tr>
<tr>
<td></td>
<td>- Extraction error</td>
</tr>
<tr>
<td>Adjustment error</td>
<td>- Same error causes than for surveys</td>
</tr>
</tbody>
</table>
## Error components and their causes

<table>
<thead>
<tr>
<th>Error components</th>
<th>Specific error causes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specification error</td>
<td>- Measuring concepts from which not enough data is available</td>
</tr>
<tr>
<td></td>
<td>- Inferring attitudes</td>
</tr>
<tr>
<td></td>
<td>- Defining valid information</td>
</tr>
<tr>
<td>Measurement error</td>
<td>- Non-trackable target</td>
</tr>
<tr>
<td></td>
<td>- Meter not installed</td>
</tr>
<tr>
<td></td>
<td>- Uninstalling the meter</td>
</tr>
<tr>
<td></td>
<td>- New non-trackable device</td>
</tr>
<tr>
<td></td>
<td>- Technology limitations</td>
</tr>
<tr>
<td></td>
<td>- Technology errors</td>
</tr>
<tr>
<td></td>
<td>- Hidden behaviours</td>
</tr>
<tr>
<td></td>
<td>- Shared device</td>
</tr>
<tr>
<td></td>
<td>- Social desirability</td>
</tr>
<tr>
<td></td>
<td>- Extraction error</td>
</tr>
<tr>
<td>Processing error</td>
<td>- Coding error</td>
</tr>
<tr>
<td></td>
<td>- Aggregation at the domain level</td>
</tr>
<tr>
<td></td>
<td>- Data anonymization</td>
</tr>
<tr>
<td>Coverage error</td>
<td>- Non-trackable individuals</td>
</tr>
<tr>
<td>Sampling error</td>
<td>- Same error causes than for surveys</td>
</tr>
<tr>
<td>Missing data error</td>
<td>- Noncontact</td>
</tr>
<tr>
<td></td>
<td>- Non-consent</td>
</tr>
<tr>
<td></td>
<td>- Non-trackable target</td>
</tr>
<tr>
<td></td>
<td>- Meter not installed</td>
</tr>
<tr>
<td></td>
<td>- Uninstalling the meter</td>
</tr>
<tr>
<td></td>
<td>- New non-trackable device</td>
</tr>
<tr>
<td></td>
<td>- Technology limitations</td>
</tr>
<tr>
<td></td>
<td>- Technology error</td>
</tr>
<tr>
<td></td>
<td>- Hidden behaviour</td>
</tr>
<tr>
<td></td>
<td>- Social desirability</td>
</tr>
<tr>
<td></td>
<td>- Extraction error</td>
</tr>
<tr>
<td>Adjustment error</td>
<td>- Same error causes than for surveys</td>
</tr>
</tbody>
</table>

Most specific error causes on the side of measurement
# Error components and their causes

<table>
<thead>
<tr>
<th>Error components</th>
<th>Specific error causes</th>
</tr>
</thead>
</table>
| Specification error | - Measuring concepts from which not enough data is available  
|                   | - Inferring attitudes                                                                    |
|                   | - Defining valid information                                                              |
| Measurement error | - Non-trackable target                                                                   |
|                   | - Meter not installed                                                                     |
|                   | - Uninstalling the meter                                                                  |
|                   | - New non-tracked device                                                                  |
|                   | - Technology limitations                                                                  |
|                   | - Technology errors                                                                       |
|                   | - Hidden behaviours                                                                       |
|                   | - Shared device                                                                           |
|                   | - Social desirability                                                                     |
|                   | - Extraction error                                                                        |
| Processing error  | - Coding error                                                                           |
|                   | - Aggregation at the domain level                                                         |
|                   | - Data anonymization                                                                     |
| Coverage error    | - Non-trackable individuals                                                               |
| Sampling error    | - Same error causes than for surveys                                                     |
| Missing data error| - Noncontact                                                                             |
|                   | - Non-consent                                                                            |
|                   | - Non-trackable target                                                                    |
|                   | - Meter not installed                                                                     |
|                   | - Uninstalling the meter                                                                  |
|                   | - New non-tracked device                                                                  |
|                   | - Technology limitations                                                                  |
|                   | - Technology error                                                                        |
|                   | - Hidden behaviour                                                                        |
|                   | - Social desirability                                                                     |
|                   | - Extraction error                                                                        |
| Adjustment error  | - Same error causes than for surveys                                                     |

Sampling and adjustment errors have no specific error causes.
Practical recommendations

1. Clearly define what your tracked data is measuring beforehand
Practical recommendations

1. Clearly define what your tracked data is measuring beforehand

**Concept:** average hours of consumption of online political news  
**Measure:** average time recorded of the visits to online political outlets’ URLs.
1. Clearly define what your tracked data is measuring beforehand

**Concept:** average hours of consumption of online political news

**Measure:** average time recorded of the visits to online political outlets’ URLs.

- What is considered a visit?
Practical recommendations

1. Clearly define what your tracked data is measuring beforehand

**Concept:** average hours of consumption of online political news
**Measure:** average time recorded of the visits to online political outlets’ URLs.

- What is considered a visit?
- Which online outlets?
1. Clearly define what your tracked data is measuring beforehand

**Concept:** average hours of consumption of online political news

**Measure:** average time recorded of the visits to online political outlets’ URLs.

- What is considered a visit?
- Which online outlets?
- Which URLs should be considered political?
1. Clearly define what your tracked data is measuring beforehand

**Concept:** average hours of consumption of online political news

**Measure:** average time recorded of the visits to online political outlets’ URLs.

- What is considered a visit?
- Which online outlets?
- Which URLs should be considered political?
- What time frame to use to compute an average?
2. Consider the impact of the chosen technologies on data quality
Practical recommendations

2. Consider the impact of the chosen technologies on data quality

**Apps**
- Where? Device
- Devices: Not iOS
- Continuous? Yes
- Types of data: URLs, Time, Device, Search terms, Incognito

**Plug-in A**
- Where? Browser
- Devices: Only PC & MAC
- Continuous? Yes
- Types of data: URLs, Time, Device, Search terms, Incognito, HTML

**Plug-in B**
- Where? Browser
- Devices: Only PC & MAC
- Continuous? No
- Types of data: URLs, Time, Device

**Proxy**
- Where? Network
- Devices: All
- Continuous? Yes
- Types of data: URLs, Time, Device
Practical recommendations

2. Consider the impact of the chosen technologies on data quality

### Apps

**Where?**
- Device

**Devices**
- Not iOS

**Continuous?**
- Yes

**Types of data**
- URLs, Time, Device, Search terms, Incognito

[Diagram of people representing different groups and categories]

iOS users ≠ Non-trackable
Practical recommendations

2. Consider the impact of the chosen technologies on data quality

<table>
<thead>
<tr>
<th>Apps</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Where?</td>
<td>Device</td>
</tr>
<tr>
<td>Devices</td>
<td>Not iOS</td>
</tr>
<tr>
<td>Continuous?</td>
<td>Yes</td>
</tr>
<tr>
<td>Types of data</td>
<td>URLs, Time, Device, Search terms, Incognito</td>
</tr>
</tbody>
</table>

- iOS users = Non-trackable
3. Explore strategies to increase the willingness of individuals to install the meter in all targets
3. Explore strategies to increase the willingness of individuals to install the meter in all targets
Practical recommendations

3. Explore strategies to increase the willingness of individuals to install the meter in all targets

- Multiple tracking technologies might need to be installed for the same participant.
- Tracking technologies present different installations processes.
- Targets (devices / browsers / networks used) are unknown.
Practical recommendations

What if we fail to properly address recommendations 2 & 3?
Practical recommendations

What if we fail to properly address recommendations 2 & 3?    Undercoverage
Practical recommendations

What if we fail to properly address recommendations 2 & 3?  

Undercoverage

Different levels of undercoverage.

- **Device:** at least one device used by a participant is not tracked
- **Browser:** at least one web-browser used by a participant is not tracked
- **In-app:** the behaviours happening inside apps are not tracked.
- **Network:** at least one network from which a participant connect to the Internet is not tracked

Undercoverage can prevent tracking the complete online behavior
Practical recommendations

What if we fail to properly address recommendations 2 & 3? → Undercoverage
Practical recommendations

What if we fail to properly address recommendations 2 & 3? → Undercoverage
Practical recommendations

What if we fail to properly address recommendations 2 & 3? Undercoverage

RESULTS

Practical recommendations

Undercoverage might be present, so what?
Practical recommendations

Undercoverage might be present, so what?
RESULTS

Practical recommendations

Undercoverage might be present, so what?
Practical recommendations

Undercoverage might be present, so what?
Practical recommendations

Undercoverage might be present, so what?

RESULTS
Practical recommendations

Undercoverage might be present, so what?

How common is this?

To what extent is this biasing results?
RESULTS

Practical recommendations

4. Define strategies to maximise the information available to identify missing data

This is still not very clear at the moment. However...
Practical recommendations

4. Define strategies to maximise the information available to identify missing data

This is still not very clear at the moment. However... we can combine survey & paradata

During the last 15 days, from how many of these different types of devices have you accessed the Internet (including using apps like Facebook, Twitter or YouTube)? Please, type the number of devices in the respective boxes.

- Computer with Windows operating system: [NUMERIC OPEN BOX]
- Apple computer(s) (MAC): [NUMERIC OPEN BOX]
- Smartphone or tablet with Android operating system: [NUMERIC OPEN BOX]
- Apple smartphone or tablet (iPhone or iPad): [NUMERIC OPEN BOX]
- Others: [NUMERIC OPEN BOX] (IF >0: “Please, specify: [OPEN TEXT BOX]”)

During the last 15 days, have you used any of the following web browsers to access the internet through a computer with Windows operating system?

<table>
<thead>
<tr>
<th>Internet Explorer</th>
<th>Chrome</th>
<th>Firefox</th>
<th>Edge, Opera or others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

During the last 15 days, have you used any of the following web browsers to access the internet through an Apple computer (MAC)?

<table>
<thead>
<tr>
<th>Internet Explorer</th>
<th>Chrome</th>
<th>Safari</th>
<th>Firefox</th>
<th>Edge, Opera or others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

During the last 15 days, have you used any of the following web browsers to access the internet through smartphone or tablet with Android operating system?

<table>
<thead>
<tr>
<th>Chrome</th>
<th>Samsung browser</th>
<th>Firefox</th>
<th>Edge, Opera or others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

During the last 15 days, have you used any of the following web browsers to access the internet through smartphone or tablet (iPhone or iPad)?

<table>
<thead>
<tr>
<th>Chrome</th>
<th>Safari</th>
<th>Firefox</th>
<th>Edge, Opera or others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

During the last 15 days, have you used any of the following web browsers to access the internet through others?

<table>
<thead>
<tr>
<th>Chrome</th>
<th>Safari</th>
<th>Firefox</th>
<th>Edge, Opera or others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Practical recommendations

4. Define strategies to maximise the information available to identify missing data

This is still not very clear at the moment. However... we can combine survey & paradata

RESULTS

• Completely covered = high chance
  true 0.

• Partially covered = not clear yet
4. Define strategies to maximise the information available to identify missing data

This is still not very clear at the moment. However... we can combine survey & paradata

---

### RESULTS

Practical recommendations

---

| During the last 15 days, have you used another device or browser apart from [INSERT DEVICE(S)] to visit the following web pages or apps: |
|---|---|
| **Yes** | **No** |
| Twitter | [ ] | [ ] |
| Facebook | [ ] | [ ] |
| The Guardian | [ ] | [ ] |
| BBC | [ ] | [ ] |
| CNN | [ ] | [ ] |
4. Define strategies to maximise the information available to identify missing data

This is still not very clear at the moment. However... we can combine survey & paradata

If the person did not use another device or browser to visit the pages/apps of interest -> undercoverage does not affect this measure

If the person did use another device or browser to visit the pages/apps of interest -> undercoverage affects this measure

Most likely cannot be done for every web page/app of interest
CONCLUSIONS

Limits

1. One specific definition of data quality.
2. Lack of previous empirical research.
3. Tracking technologies are constantly evolving.
4. Metered data errors are considered independently.

Take-home messages

1. Using metered data is complex and many decisions must be taken.
2. Reporting these decisions and conducting robustness checks is necessary.
3. More empirical research is needed.
4. This framework can help on all these aspects.
5. Identifying when a lack of behaviour is real or a product of undercoverage is key.
6. Confounding both phenomena can inflate measurement and missing data errors.
Thanks!

Questions?

Oriol J. Bosch

- o.bosch-jover@lse.ac.uk
- orioljbosch
- https://orioljbosch.com/