Data Linking I:

Survey data & social media data

Johannes Breuer,
GESIS Data Archive

CESSDA Training Day
27-28 November 2019, Cologne

cessda.eu @CESSDA_Data

cessda

gesis Leibniz Institute for the Social Sciences

What is social media data?

- Simple definition: Any data generated by users of social media platforms, such as Facebook, Twitter, YouTube or Reddit
- More specific definition difficult due to differences between platforms:
  - Different types of interactions/use
  - Different types of data

gesis Leibniz Institute for the Social Sciences

Examples of social media data

- Facebook
  - Posts & comments
  - Photos
  - Profile information
- Twitter
  - Tweets, retweets, & replies
  - Profile information
- YouTube
  - Video or channel statistics
  - Viewer comments

Why collect social media data?

- Social media use have become ingrained in everyday life for many people
- This use generates a lot of data
- A lot of these data are also interesting for research in the social sciences

Social media data in the social sciences

- Most studies/researchers use text data
- Other types of social media data (e.g., photos, videos) less commonly used
- Unit(s) of analysis for social sciences typically are the user(s)

How can you collect social media data?

- Basically 3 ways (Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2019). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*, Accepted for publication):
  1. Collect data yourself
  2. Direct cooperation with social media company
  3. Buy data from data reseller or market research company
- Choice should essentially depend on the research question and the available resources (knowledge/skills, time, money…)

Collect your own data (CYOD)

- Again, 3 options:
  1. Application Programming Interfaces (APIs) of platforms
  2. Web scraping
  3. “Data donation”
- each option has specific advantages and disadvantages

APIs

Advantages	Disadvantages
• Most social media platforms provide them • Usually good documentation • Many software tools and packages (e.g., for R and Python) available for collecting social media data via APIs • Provide structured data (often in the form of JSON files)	• APIs typically have rate limits for requests (that can also change substantially) • Can be limited, changed or even closed off entirely (prime example: Facebook essentially closing its Graph API in the wake of the Cambridge Analytica scandal)

APIs

- Freelon (2018): Computational Research in the **Post-API Age**
  - argues for the importance of moving (back) to web scraping in computational research
- interesting exchange between Bruns (2019) and Puschmann (2019) about consequences of changes/closing of APIs recently published in *Information, Communication, & Society*

Web scraping

Advantages	Disadvantages
• Flexible • Does not depend on goodwill of social media companies	• More complicated/involved (than access via APIs) • Changes in website structures can be an issue • Data has to be structured/cleaned • Can be (deemed) a legal grey zone

Data donation

- people can download their personal data from most platforms (in many cases implemented in reaction to the new European GDPR) which they could then share with researchers
- Halavais (2019) proposes this as a solution for "overcoming terms of service"
- Thorson et al. (2019) used this approach to study "exposure to news and politics on Facebook" (with a student sample)
- browser plugins can be another option (see Haim & Nienierza, 2019 for an example for Facebook data)

gesis Leibniz Institute for the Social Sciences

Data donation

Advantages	Disadvantages
• Informed consent • Transparency for the users • No issues with rate limits, terms of service, etc.	• Not easy to implement (users have to be instructed, data has to be safely uploaded…) • Solutions for anonymization required (example: friends tagged in participants' Facebook posts)

What is data linking?

- Combining data from different sources for the same units of analysis (e.g., individuals)
  - in the quantitative social sciences usually survey data + X
- Different terms in the literature:
  - Data linking
  - Data linkage
  - Record linkage
- 2 basic linkage/linking types:
  1. Deterministic
  2. Probabilistic
- focus here on deterministic linkage or linking: unique identifiers (or combination of identifiers) allows direct matching of units of analysis

Why link surveys & social media data?

- Self-reports can be biased
  - social desirability
  - problems with recall
- Social media data alone can be difficult to use as they tend to lack…
  - information about the individuals being studied (e.g., attitudes, personality…)
  - relevant outcome variables (e.g., voting intention/behavior)
  - explicit informed consent
- Linking to alleviate limitations of the two data types (Stier et al., 2019)

How to link surveys & social media?

- 4 general types of linking

a) When does the linking happen?
   1. Ex ante: Data are collected together (for the same time period)
   2. Ex post: Data that have been collected are linked with existing data

b) On what level are the data linked?
   1. Individual level
   2. Aggregate level

Use cases

- Methodological questions, e.g., regarding over- oder underreporting in surveys: Haenschen (2019) used a combination of survey and Facebook data to measure political activity on the platform
- Political attitudes and opinions: Pasek et al. (2019) use data from polls and Twitter to assess attitudes towards US presidents
- Many other applications possible: e.g., to study media use, social networks or well-being

Challenges

- Working with linked survey and social media data creates specific challenges for all phases of the research data lifecycle
- Exemplary key issues:
  - Recruitment of participants
  - Informed consent
  - Privacy & data protection

Recruiting participants

- Two options:
  1. Collect social media data & recruit people via that social media platform
  2. Collect survey data, then ask for consent to collect social media data
- How you recruit your participants and what method you use to collect the social media data affects the composition of your sample (and might introduce different biases)
- Choice should depend on your research question

gesis Leibniz Institute for the Social Sciences

Informed Consent

- Important to collect informed consent (esp. in Europe with GDPR) when you link surveys and social media data
- Make clear what data you collect, why you collect it and how it will be used and stored (also mention data sharing if applicable)
- Al Baghal et al. (2019) provide a good template (they linked surveys and Twitter data): short informed consent with important basic information in the survey + extended (privacy) information that is optional to read for participants

gesis Leibniz Institute for the Social Sciences

Privacy & data protection

- Need to pay special attention to the privacy of people who did not consent to their data being used in the study (e.g., Facebook friends tagged in posts)
- Some of the resources for sharing social media data can also be used as guidance for dealing with linked survey and social media data (e.g., Bishop & Gray, 2017; Kinder-Kurlanda et al., 2017; Mannheimer & Hull, 2017; Williams et al., 2017)

gesis Leibniz Institute for the Social Sciences

Case study

- internal GESIS research project with aim of studying use of online media (esp. news)
- Methods
  - Web tracking panel from market research company
    - ~ 2000 participants per month
    - data for one year
    - ~ 94 mio. data points (visits: domain level)
  - Additional data for parts of the sample
    - Tracking of mobile app use
    - Data from 3 online surveys (focus: media use and politics)
    - Social media data: Twitter, Facebook, Spotify

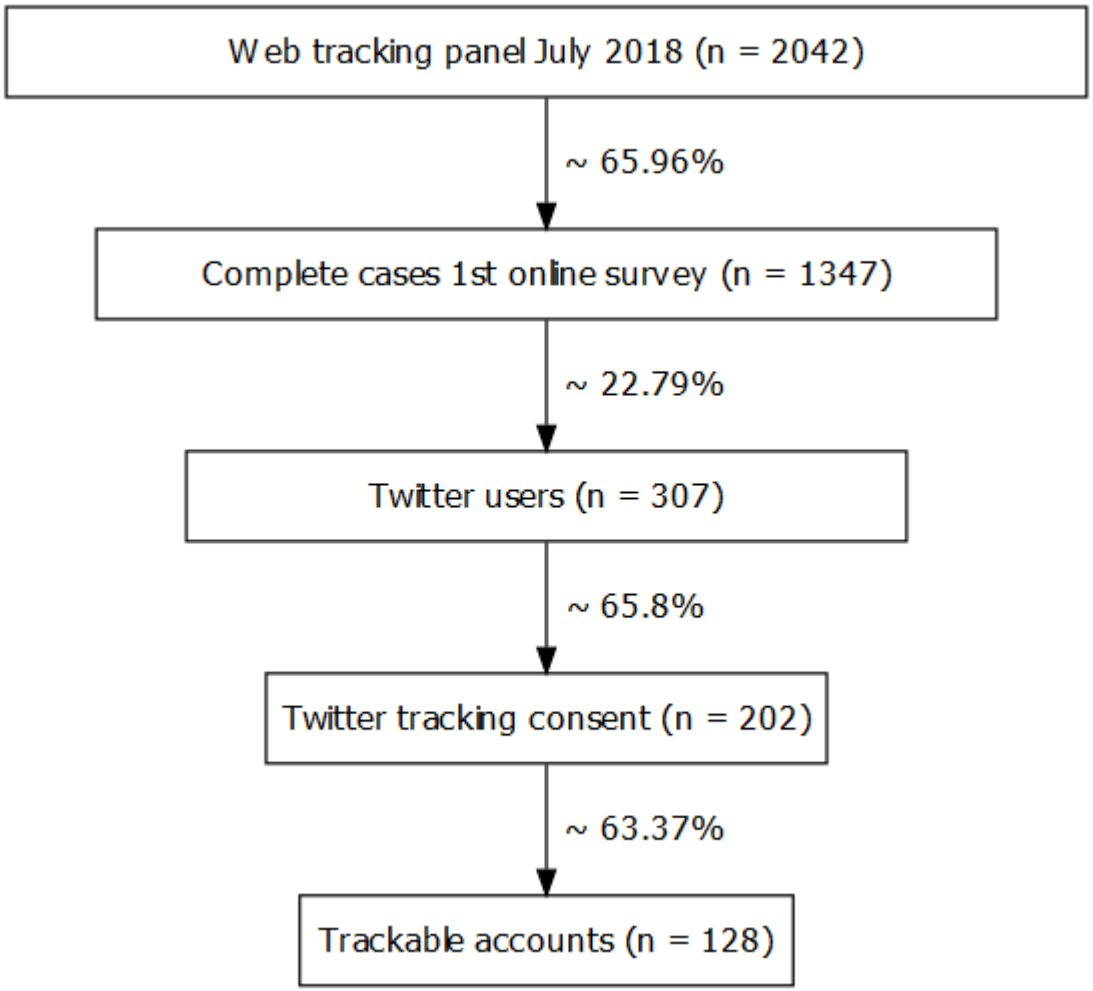
gesis Leibniz Institute for the Social Sciences

Social media data in our project

- consent collected via online surveys
- Twitter
  - Continuous tracking using public streaming API
- Facebook
  - Browser plugin (Haim & Nienierza, 2019)
    - for Firefox and Chrome
    - collects public posts (+ some metadata) from users‘ feeds
- Spotify
  - Web app developed at KU Leuven
  - Collects 50 most recently played songs, playlists, and preferences

gesis Leibniz Institute for the Social Sciences

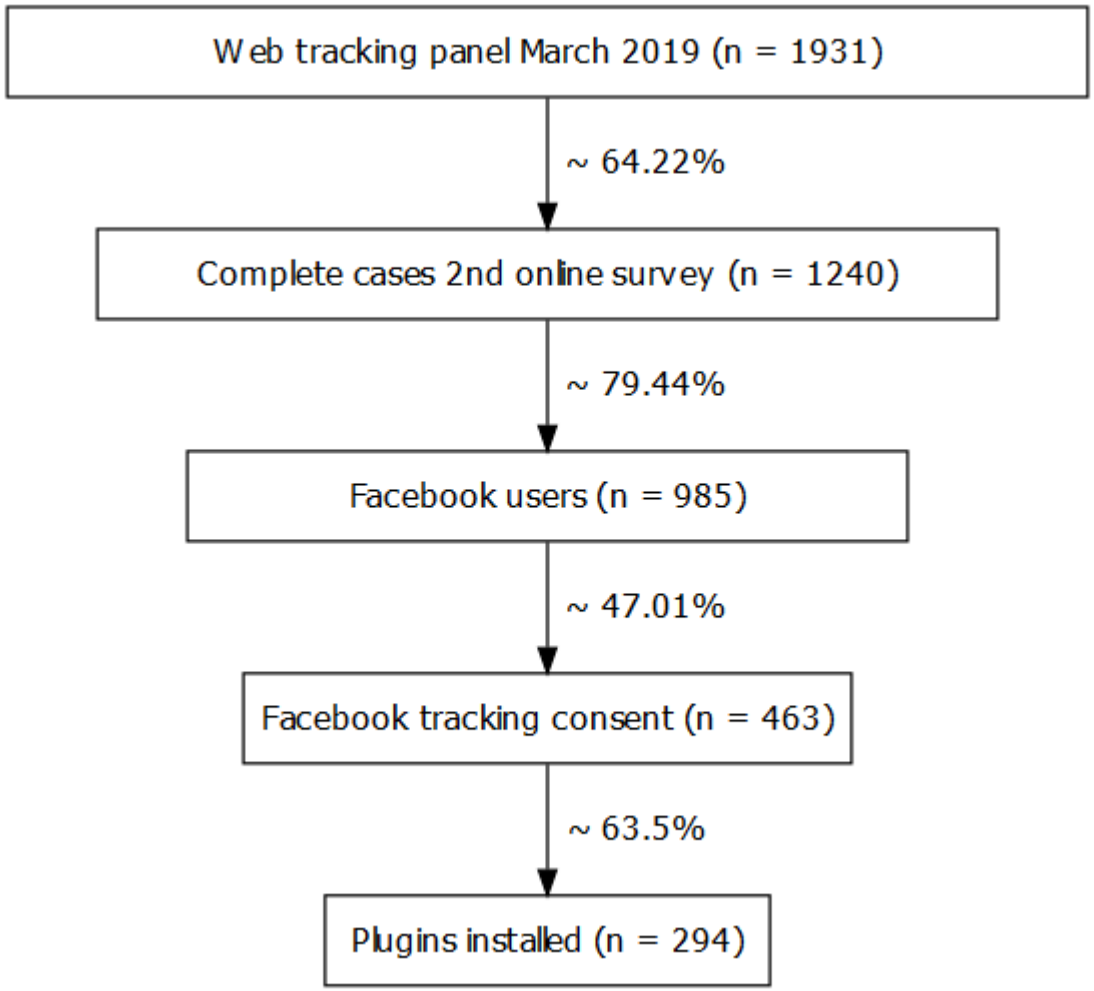
Twitter data



Member of the Leibniz Association

gesis Leibniz Institute for the Social Sciences

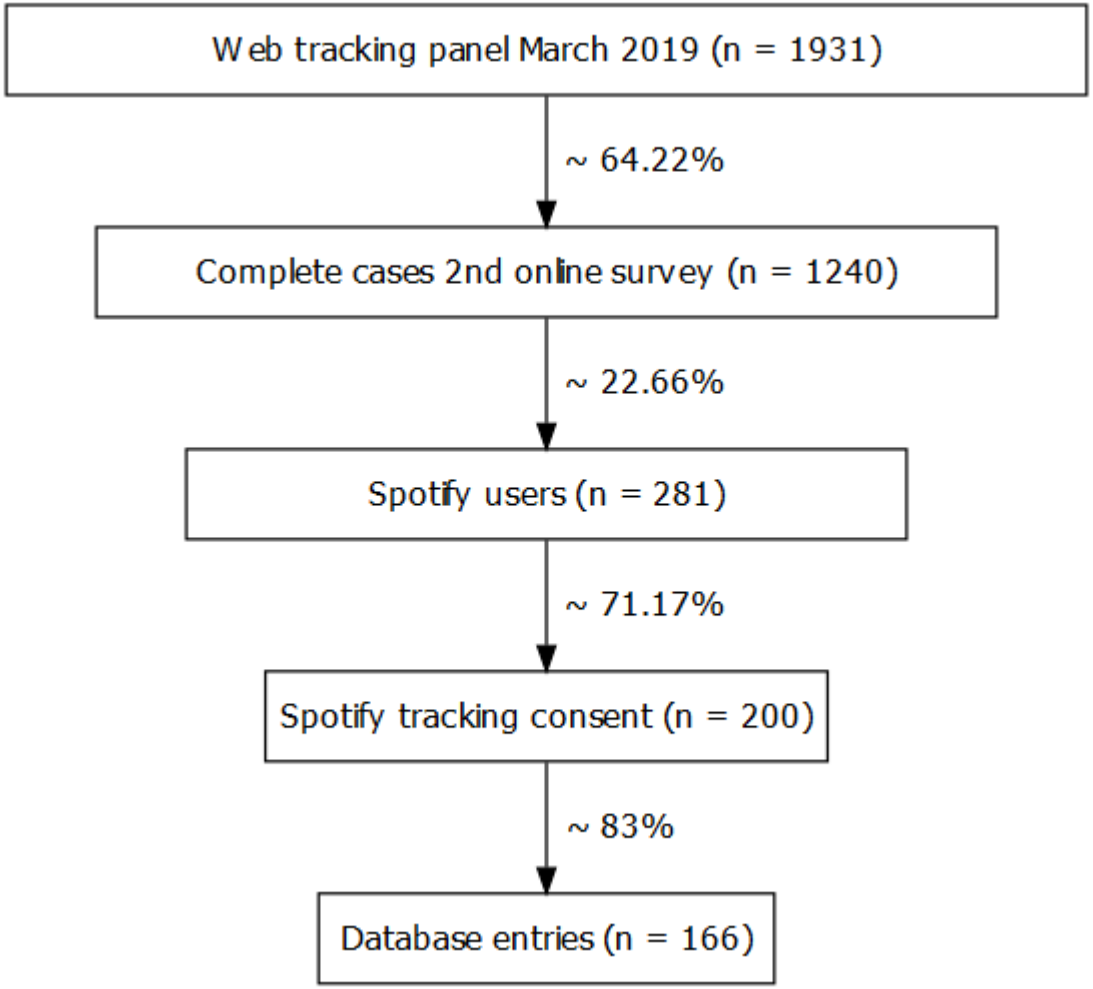
Facebook data



Member of the Leibniz Association

gesis Leibniz Institute for the Social Sciences

Spotify data



Member of the Leibniz Association

gesis Leibniz Institute for the Social Sciences

Next data processing steps in our project

- Check for systematic bias in the dropout stages
- Check data quality
  - Quantity of user activity: data points per person
  - Quality of user activity: e.g., active vs. passive Twitter use or importance of Facebook as news source
- Find solutions for making the data available
  - Full raw data cannot be shared (example: people frequently visiting their personal homepage)

Member of the Leibniz Association

Pros and cons of our approach

Advantages	Disadvantages
• Individual-level data • Large and heterogeneous sample • Large bandwidth of data • Informed consent from participants • Easy access to web tracking data • Facebook data without need to use API	• Potential biases in the sample • High costs • Need to use APIs for Twitter & Spotify • Changes to Facebook feed structure can be problematic for browser plugin

Conclusion

- There are different ways of collecting social media data, each with their own pros and cons
- Linking surveys and social media data can help in alleviating some of the limitations of these data types
- The choice of data collection and linking methods should depend on your research question (but also take into account what resources are available)
- Recruiting participants, collecting informed consent, and protecting participant privacy are some of the key challenges when working with