Zoltán Kmetty, Anna Vancsó, Dániel Váry, Ádám Stefkovics

Acquiring personal Facebook data

Is it still possible?

PROJECT WEBSITE - https://fbpilot.tatk.elte.hu/en/

The research is supported by NKFI, under the grant agreement: 12899

Principal Investigator
Zoltán Kmetty PhD
zkmetty@tatk.elte.hu

Introduction

- There were over 2.07 billion monthly active FB users in 2017 (Q3), and 1.37 billion daily active users. In every minute 510 000 comments are posted and around 300k status are updated
- In Hungary **90 percent of the active internet users** have a profile on Facebook
- Hungarians mainly use FB to **contact** friends and relatives, but half of them use it **as a news feed**, so it plays an extremely important role in their everyday life, according to their level of knowledge and way of thinking.
- Facebook is **not the main target of researchers**: a vast number of researches deal with Twitter data (Tinati et al 2014), but relatively few with FB data
    - it is easier to structure Twitter data: tweets are short and the number of possible actions a user can do is rather limited
    - it is much easier to get data from Twitter, through its API-s
- The case of FB, contents of the public sites (eg. restaurants, institutions, telco providers, universities, public figures, etc.) are possible to collect (through its Graph API), but **contents of pages of users are prohibited to gather**, even if these contents and activities are public and not private

Possible ways of (legal) data collection

1. **Public FB sites data through Graph API**
    - This data collection strategy is good for some specific research questions, for example to study the FB activity of politicians
    - Public sites contain only part of the FB traffic, and not a random part!
    - FB **restricts** now this access

2. **Through Facebook application**
    - Fill out a test and in exchange, let the researchers download the FB profiles of respondents
    - myPersonality Project as an example
        - more than 6 million test results, and more than 4 million FB profiles
        - data collection stopped at 2012
    - After Cambridge Analytica scandal, this data collection method is quasi **forbidden**

3. **Facebook gives access to data for researchers**
    - Previously only few chosen researchers have the privilege for that
    - Social Science One is a new project, with a partnership of academy and Facebook
        - Launched in 2018 April
        - Research teams have to go through a grant process
            - If their project is supported, they can get access to anonymized Facebook data
                - Crowdtangle API
                - Ad Library API
                - URL shares dataset

4. **Through web-browser plugin**
    - Installed web-browser plugins could monitor the users FB activities if consent given
    - LMU Munich plugin - https://www.fbforschung.de
    - Firefox and Chrome
    - collects public posts

5. **Users share their downloaded FB data with researchers**

    Our project follow this strategy

Our research design

Objectives

Analyze private FB data, not only posts shared on public sites
Link survey and FB data
Analyze which kind of **research questions** can be answered based on individuals FB data

Strategy

Recruit participants – 150 people (quota sample (age))

↓

Ask the participants to fill out a questionnaire (F2F)

↓

Ask them to download their Facebook data and share this data with us

Our research design

Questionnaire cover different topics

- General FB usage
- Self-representation in FB
- News consumption in FB
- Politics (political network homophily, party preferences)
- Leisure time activities
- Music preferences
- Depressions scale
- Demographic

ME1. Please rate the following statements on how characteristic they are for you.
1-7 scale of the answers:
1- not characteristic at all, 7 – very typical of me
-1 DK/NA

I feel as if I don't know myself very well.	1	2	3	4	5	6	7	-1
I feel alienated from myself.	1	2	3	4	5	6	7	-1
I don't know how I really feel inside.	1	2	3	4	5	6	7	-1
I always stand by what I believe in.	1	2	3	4	5	6	7	-1
I am true to myself in most situations.	1	2	3	4	5	6	7	-1
I think it is better to be yourself, than to be popular.	1	2	3	4	5	6	7	-1
I live in accordance with my values and beliefs.	1	2	3	4	5	6	7	-1
I usually do what other people tell me to do.	1	2	3	4	5	6	7	-1
Other people influence me greatly.	1	2	3	4	5	6	7	-1
I am strongly influenced by the opinions of others.	1	2	3	4	5	6	7	-1
I always feel I need to do what others expect me to do.	1	2	3	4	5	6	7	-1

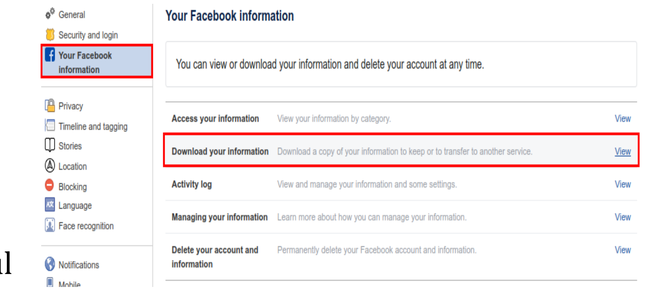
Facebook data

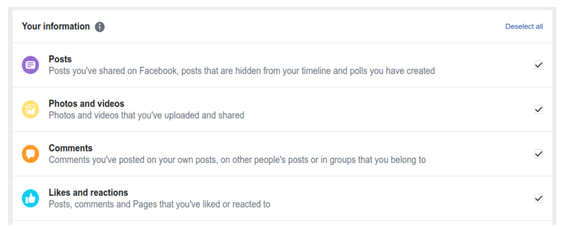
Ask participants to download their personal FB data, through the site

Some of the data tables contain very **sensitive information** (FB messenger), and some of them don't contain any useful information from research perspective, so we focus on specific tables:

- Posts
- Likes and reactions
- Friends
- Following and Followers
- Groups
- Profile information
- Ads

Data format: Json





GDPR

Participants sign **consent form** before the interview
The research team follow **strict data privacy protocol**
We can only access and use **anonymized data**
After downloading FB data in Json form, an R script is used to pre-process the data
Data format changed to CSV (small data loss, but more handy format for social scientists)
Names are masked
MD5 method (openssl package in R)
We can only mask those names which are in the contact list
Json files are deleted researchers only have access to processed files

V1	V2	V3
1560349774	@a70efd846a97bf62fdc663739bb54660 replied to [redacted] comment.	[redacted] igy még jobb!
1560320907	@a70efd846a97bf62fdc663739bb54660 replied to [redacted] comment.	@132177d8da160b147b35237b36c68181 de azért az meno, hogy az embereknek a Napló jut eszébe
1560318900	@a70efd846a97bf62fdc663739bb54660 replied to [redacted] comment.	@132177d8da160b147b35237b36c68181
1560194900	@a70efd846a97bf62fdc663739bb54660 commented on @2f8ce639df0a33896978e0f58b37763d 's video.	Én is!!!
1560107107	@a70efd846a97bf62fdc663739bb54660 commented on @b6d090d63601ef46ed38608256de609e 's photo.	Nagyon boldogat!!!
1560026157	@a70efd846a97bf62fdc663739bb54660 commented on @132177d8da160b147b35237b36c68181 's photo.	Pont itt voltunk Dariaval es Dianaval par hete! Nagyon boldog évfordulót nektek!!!
1559806685	@a70efd846a97bf62fdc663739bb54660 commented on @0963eeeb7e62f900408f02a7ee3f009a 's post.	Boldog szülinapot Lujzinak!
1559686993	@a70efd846a97bf62fdc663739bb54660 commented on [redacted] 's post.	@131fb0c89d8beaa31baeb7036b6c46b5
1559562369	@a70efd846a97bf62fdc663739bb54660 replied to @d9d70e28e5131b5c5f4b626fd5cbda97 's comment.	@d9d70e28e5131b5c5f4b626fd5cbda97 már kint vannak, igy most nem ti lesztek!

Challenges

- People names who are not in the **friend list** are not masked after the first round of anonymization
- **Nicknames or not tagged names** are not anonymized by the script
- **Human processing** needed

Limitations - fragmented pieces

Although we will have access to private FB data, this data will be **fragmented** in many ways
In the case of comments by participants we don't know the **original content**. So we see a comment, but don't know what was the original post
In the case of likes and reactions we also don't know the **original content**
We can not follow a **whole thread**, just the some pieces of it
This is a serious limit, but totally understandable from GDPR point of view

Planned papers

Wide spectrum of papers are planned:

- Methodological paper (pros and cons of this research strategy)
- Self representation and posting habits
- Depression and posting habits
- Music tastes and liked pages of musicians and bands (omnivore – univore taste)
- Politics related posts in off- and on-campaign period

Collaboration is possible, we're open for **joiners**
We can share the survey data and aggregated FB data (not raw data)