Utrecht University

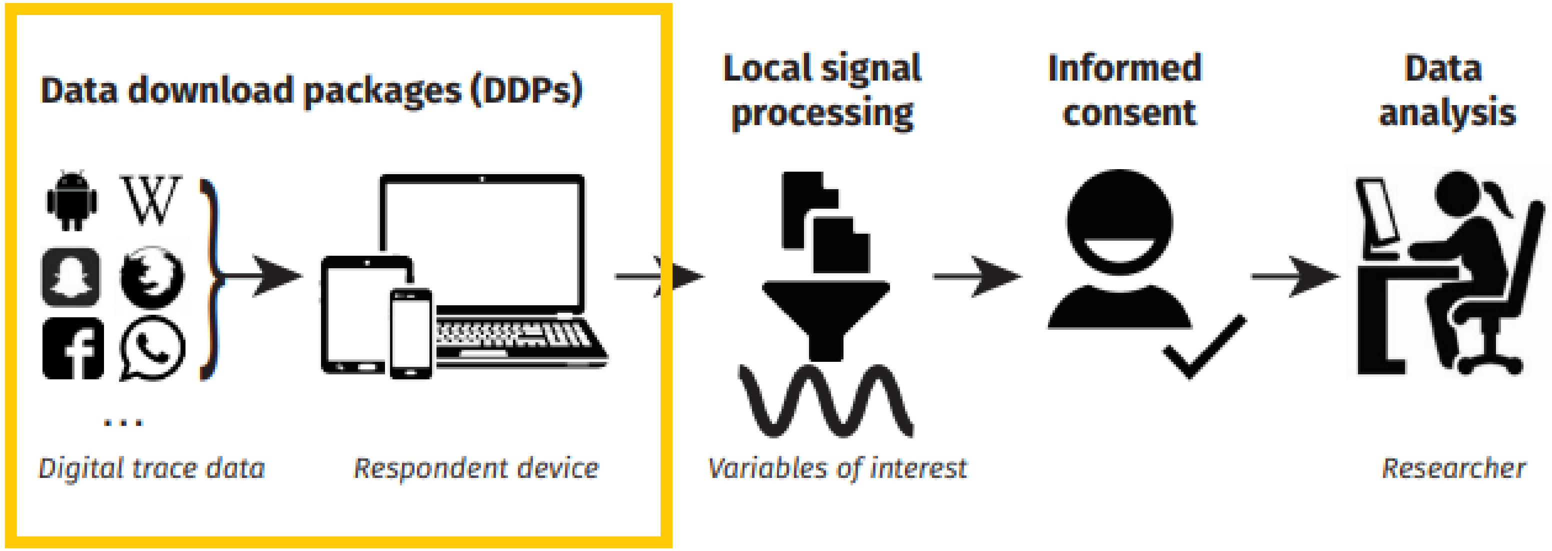
12-09-2023

Volatility of data download packages

DATA DONATION SYMPOSIUM
ZURICH, SEPTEMBER 2023

Thijs Carrière[1], Laura Boeschoten[1], & Niek de Schipper[2]
[1]Utrecht University, [2]University of Amsterdam

Introduction: Data donation through DDPs



- DDPs are a central element of data donation approaches.

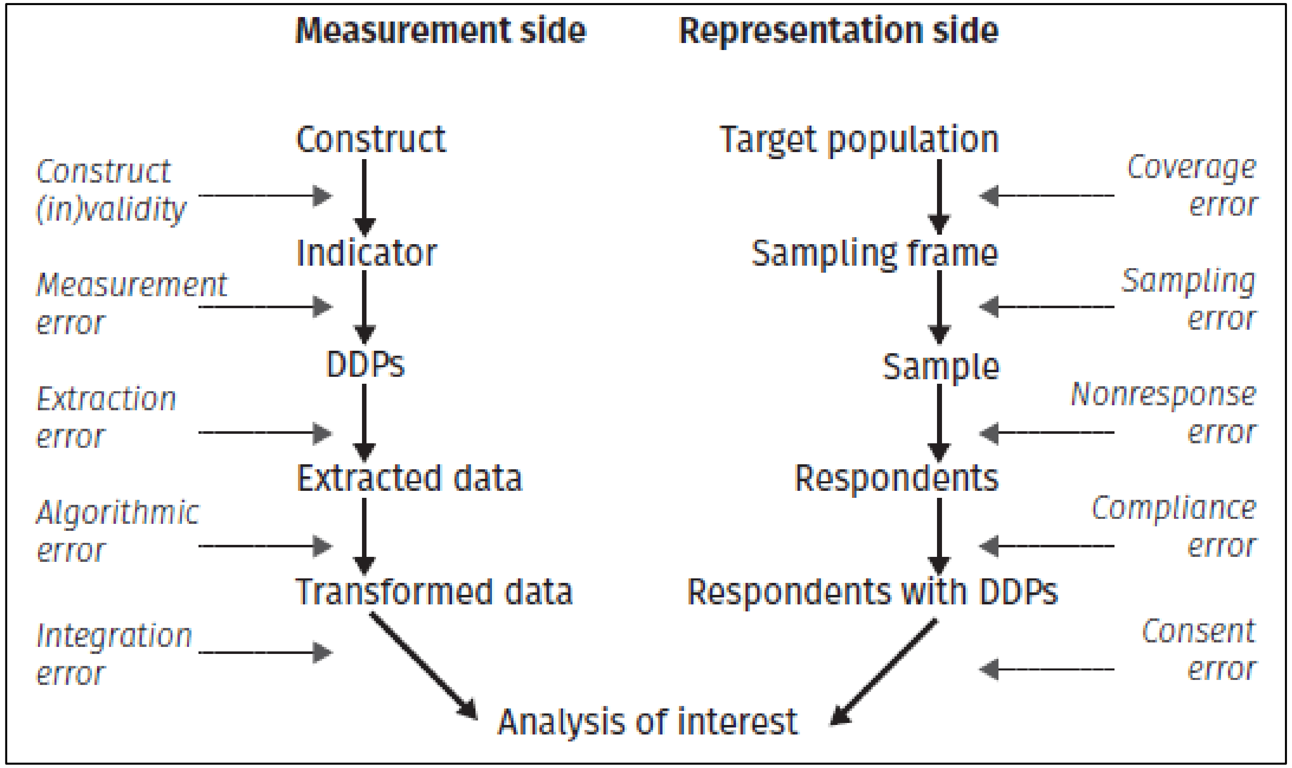
- Change occurs in platform features & DDPs:

(e.g. WhatsApp data donation by Corten et al.; Facebook for Silber et al., 2022)

Introduction: Effect of change in DDPs

- Change in DDPs threat for data quality.

- TE framework (Boeschoten et al., 2022) summarizes error sources.



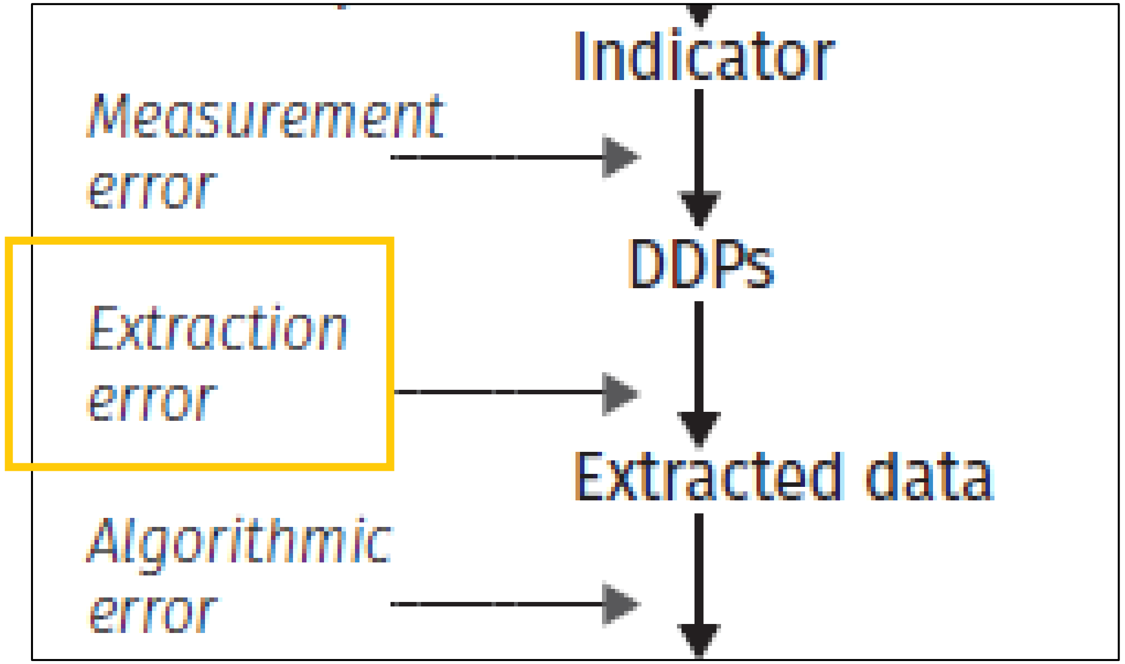
Extraction error

- Error in extracting data from DDPs.

  (e.g. not extracting all data or extracting incorrect data)

- Change in DDPs makes extraction error more likely.

- Robust scripts could account for change and extraction error.

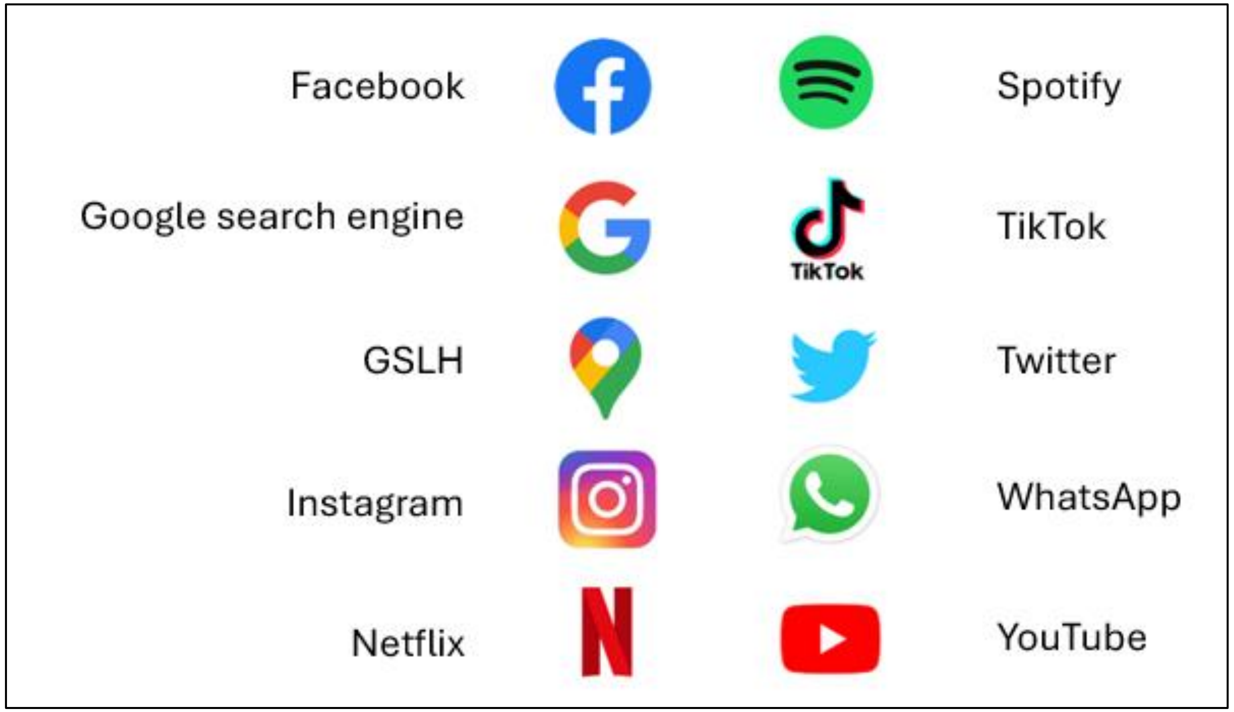


Goals of the study

- Creating an overview of change over time in structure and content of DDPs.
- Comparison of operating systems on their change over time differs (Apple and Android).
- Working towards recommendations in building extraction scripts.

The study

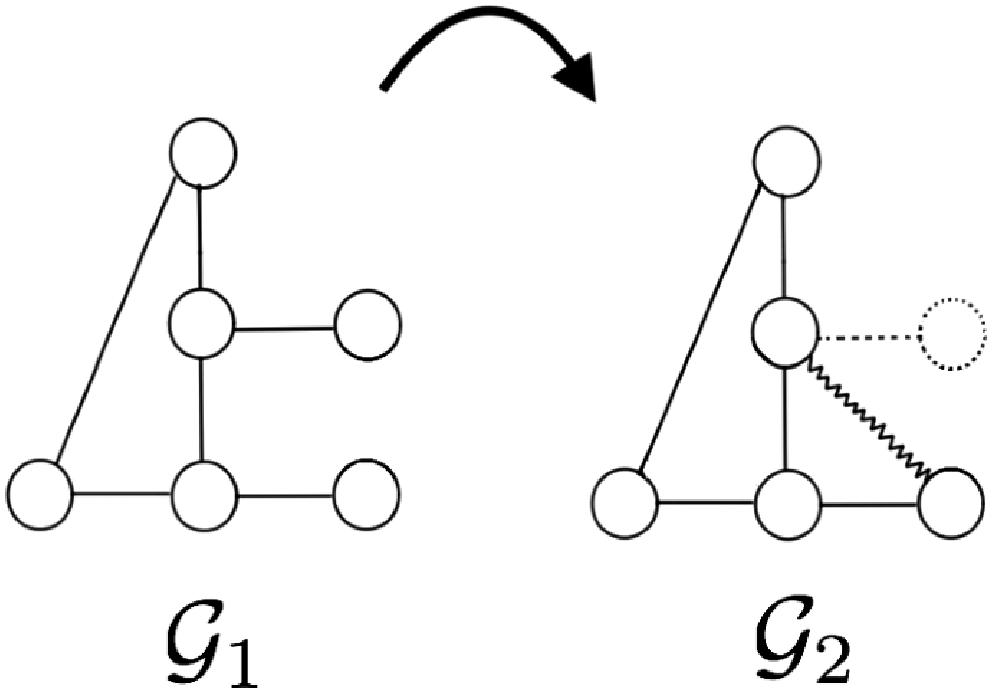
- 10 platforms
- 2 fake accounts (Apple and Android)
- Systematically use of platforms + collecting DDPs
- Period of 5 months
  (9 – 15 DDPs per account)



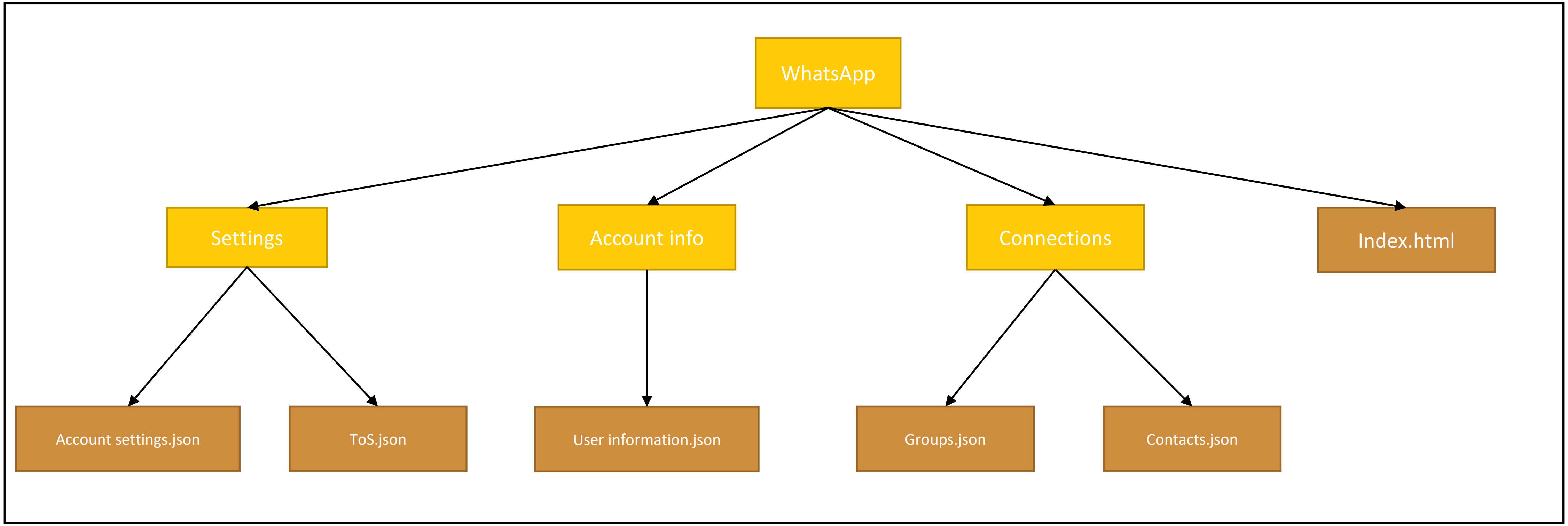
DDPs were collected on these 10 online platforms

Analysis plan

- Metric for change in networks: Graph Edit Distance (GED)
- GED = sum of change in nodes and edges.
- Folder structures can be seen as tree-structured network graphs.



Tree graph of folder structure



GED for DDP change

GED for DDP change needs some adjustments

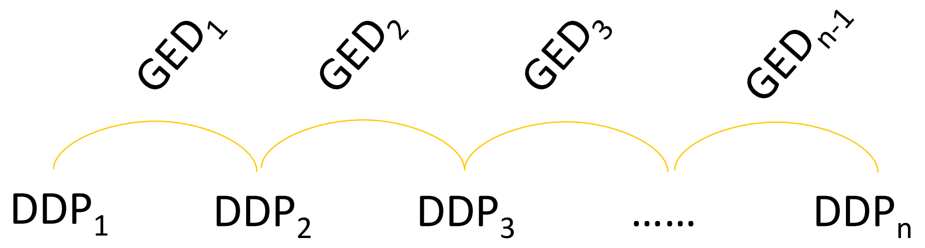
GED for DDP change:

- New nodes are no problem
- Every node only has 1 edge upwards;

  → any change in edges is relocation (should be counted once)
- GED better comparable over platforms when relative (optional)

Our GED for DDP change:

- $$\frac{Deleted\ Nodes\ +\ Relocated\ edges}{Total\ number\ of\ nodes\ in\ DDP1}$$
- $Deleted\ Nodes\ +\ Relocated\ edges$

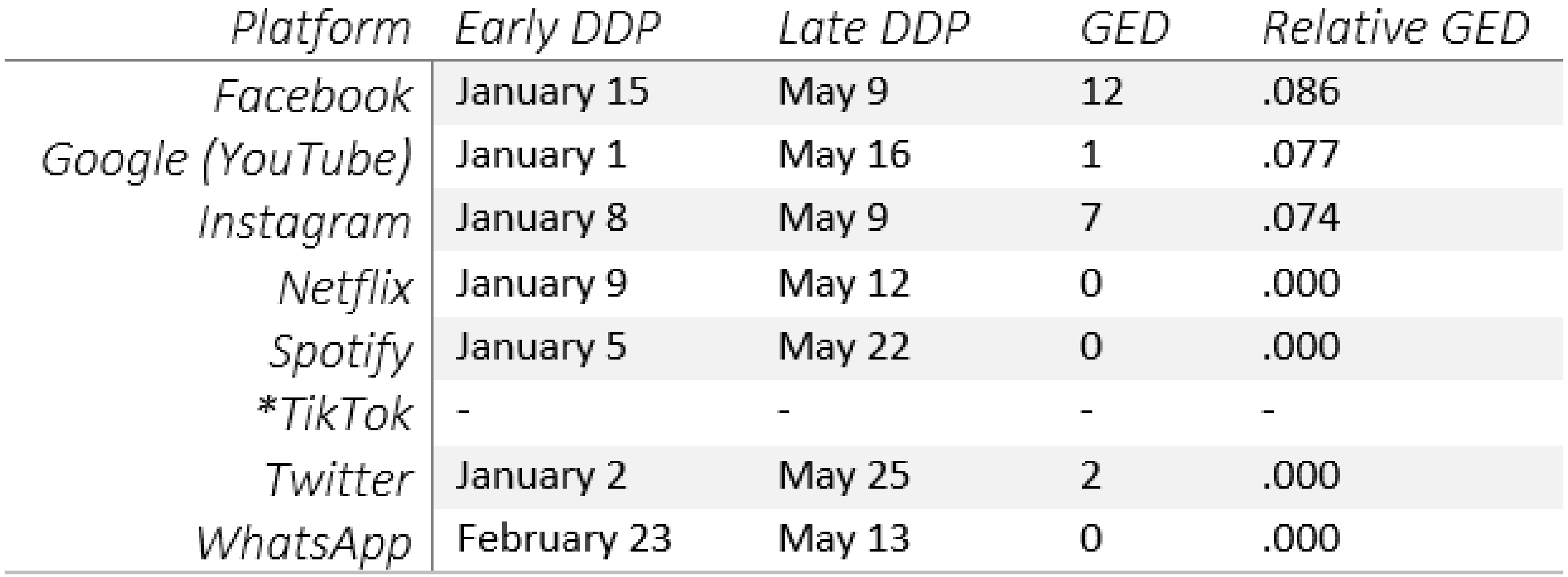


GED_1 GED_2 GED_3 GED_{n-1}

DDP_1 DDP_2 DDP_3 …… DDP_n

+ qualitative check of changed files

Folder structure: overall results



Platform	Early DDP	Late DDP	GED	Relative GED
Facebook	January 15	May 9	12	.086
Google (YouTube)	January 1	May 16	1	.077
Instagram	January 8	May 9	7	.074
Netflix	January 9	May 12	0	.000
Spotify	January 5	May 22	0	.000
*TikTok	-	-	-	-
Twitter	January 2	May 25	2	.000
WhatsApp	February 23	May 13	0	.000

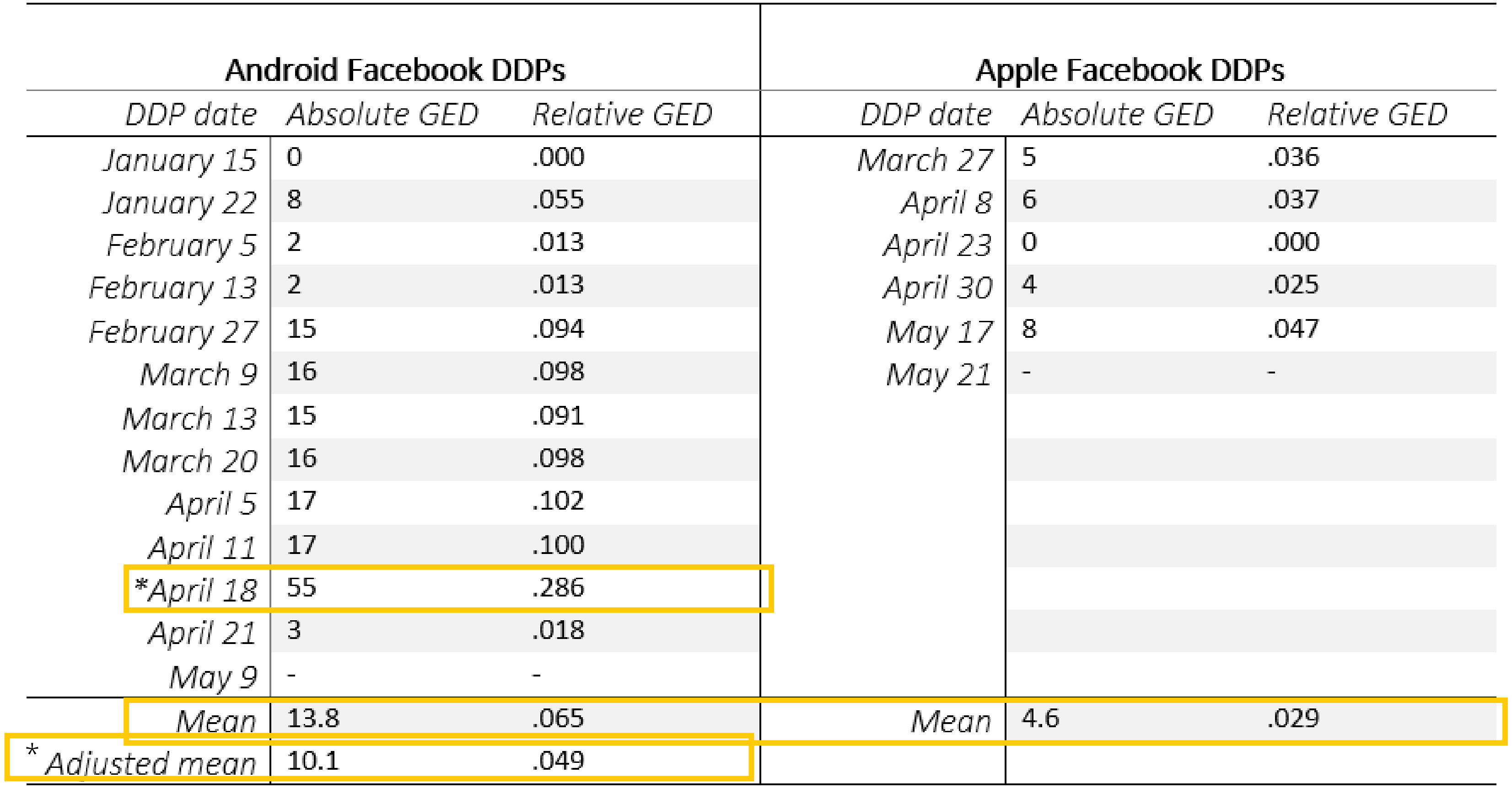
**DDP consists of single file*

More illustrative GED results for Netflix and Facebook.

Folder structure results

- Netflix had no change in the DDP structure at any point in time (mean GED = 0)

- Facebook:
  - Language of files/ filenames changes (Android)
  - Minor files disappear/ reappear
  - Names of files change regularly



Android Facebook DDPs			Apple Facebook DDPs		
DDP date	*Absolute GED*	*Relative GED*	*DDP date*	*Absolute GED*	*Relative GED*
January 15	0	.000	*March 27*	5	.036
January 22	8	.055	*April 8*	6	.037
February 5	2	.013	*April 23*	0	.000
February 13	2	.013	*April 30*	4	.025
February 27	15	.094	*May 17*	8	.047
March 9	16	.098	*May 21*	-	-
March 13	15	.091			
March 20	16	.098			
April 5	17	.102			
April 11	17	.100			
**April 18*	55	.286			
April 21	3	.018			
May 9	-	-			
Mean	13.8	.065	*Mean*	4.6	.029
* *Adjusted mean*	10.1	.049			

Concluding remarks

- Change in structure and content occur to varying degrees.
- DDP change needs to be considered to account for extraction error.
- Facebook and Instagram seem relatively unstable in both structure and content
- Hard to phrase good generalizable recommendations
  → Platforms differ much
  → Change is not stable
  → Language seems a common problem

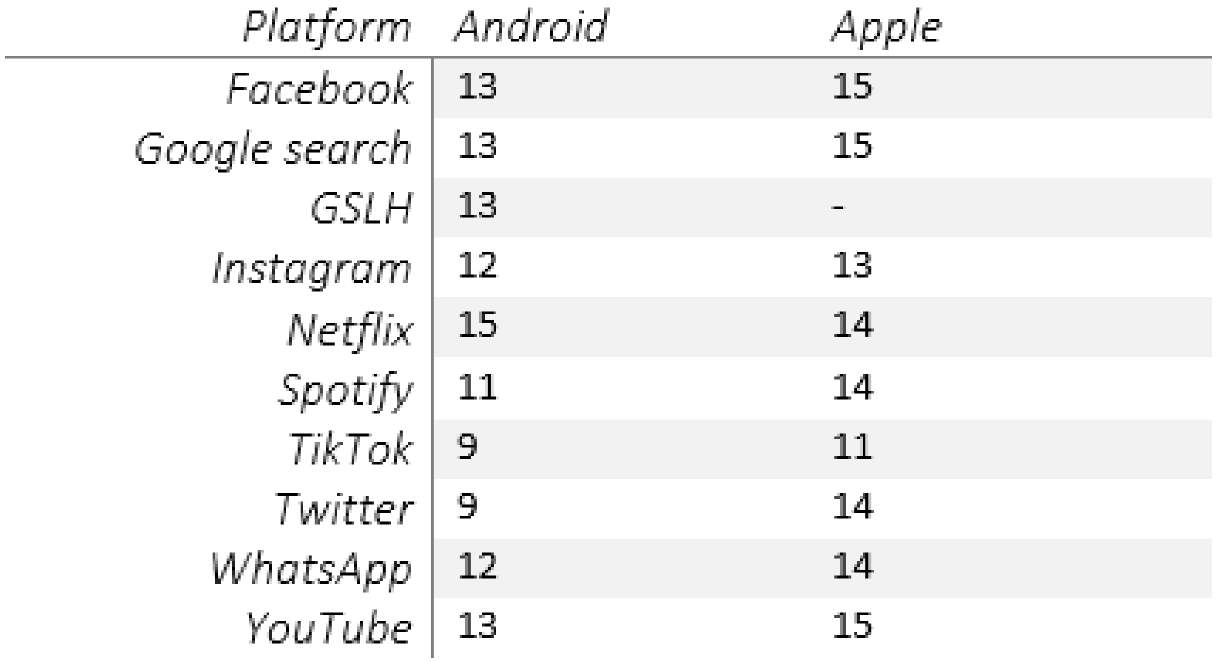
Utrecht University

Sharing science,
shaping tomorrow

Thank you for your attention!

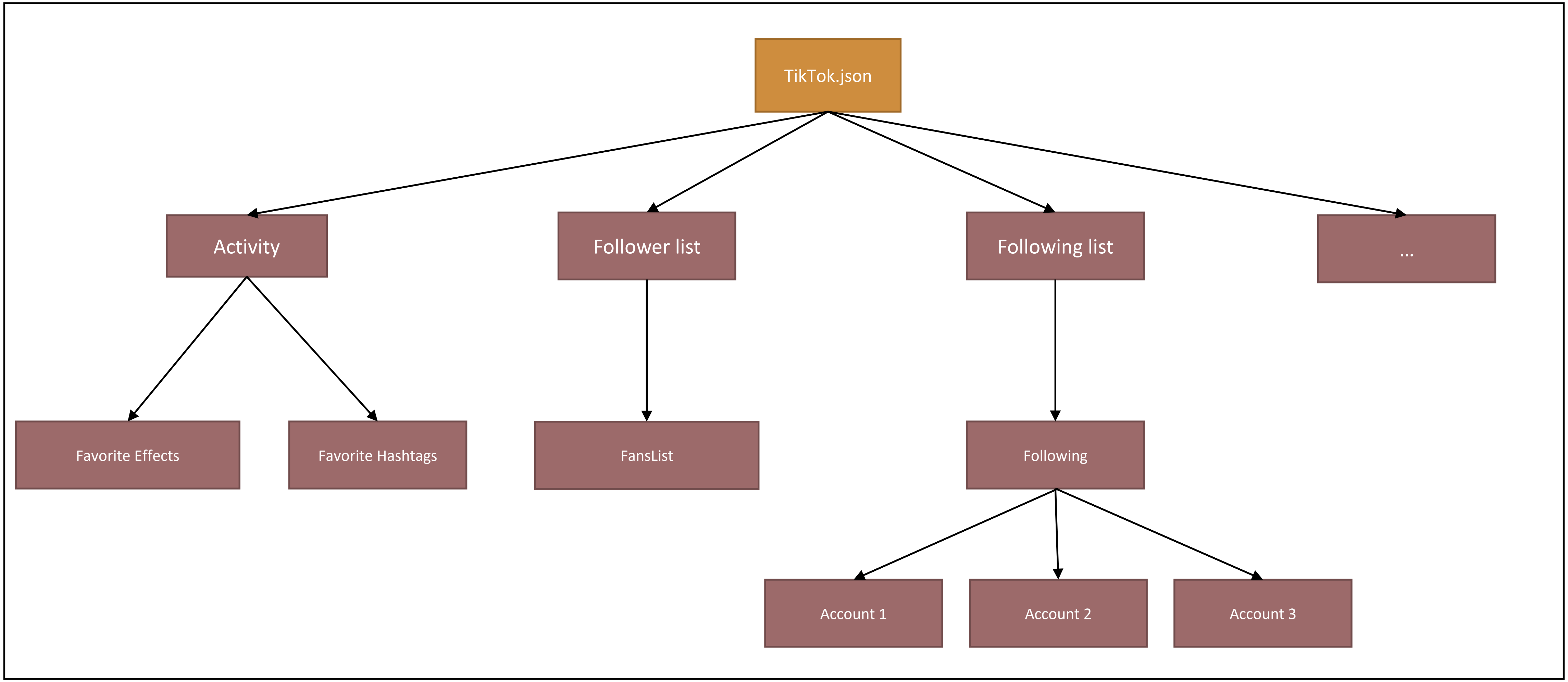
t.c.carriere@uu.nl

Results: The collected data



Platform	*Android*	*Apple*
Facebook	13	15
Google search	13	15
GSLH	13	-
Instagram	12	13
Netflix	15	14
Spotify	11	14
TikTok	9	11
Twitter	9	14
WhatsApp	12	14
YouTube	13	15

Tree graph of folder structure



JSON structure results

- TikTok (JSON is full DDP)
  → Shows GEDs of 0.
  → Values (content) does change: (urls)

- Facebook (Android)
  → 'Your posts'-json.
  → GED occurences of .039/.096;
  → Values change in language.

- Facebook (Apple)
  → 'Your Posts"-json.
  → Single GED occurence of .006;
  → Deleted post changes structure.