

Going beyond survey self-reports: Processing, enriching, and analyzing digital traces

Claassen¹, Boeschoten², Nguyen², Struminskaya², & van Es²

¹DZHW, Leibniz University Hannover

²Utrecht University

CIPHER 2026

Washington, DC (USA) – February 25 to 27, 2026

Introduction

- Societal trends, including digitalization, are primarily studied with self-reports
 - *Self-reports struggle with recall error and social desirability bias* (Baghal et al., 2014; Kreuter et al., 2008)
 - *Discrepancies between passively measured and self-reported digital media use* (Parry et al., 2021)
- Online panels have begun collecting digital traces to complement self-reports
 - *For example, this includes the GESIS Panel.dbd and LISS Panel*
 - *Traces are collected based on various approaches, including data donation and web tracking*
- Methodological research has mostly focused on collecting traces
 - *Traces usually lack important context information* (Wedel et al., 2024)
 - *Various analytical decisions may influence research outcomes* (Ochoa & Revilla, 2025)
 - *Thus: Demand for standardized workflows to process, enrich, and analyze traces*

Case Study and Research Question

- **Case study**: Measuring digital streaming behavior in the context of Netflix
 - *Netflix has more than 260 million subscribers in over 190 countries* (Netflix, 2024)
 - *Video streaming services foster new viewing habits* (Tana et al., 2019; Kaur & Ashfaq, 2023)
 - *This digitalization trend cannot be studied with self-reports alone* (Lobato & van Es, 2025)
- **RQ**: How can digital traces from Netflix users be transformed into individual-level genre measurements?

Method: Trace Collection and Workflow

- Data donation (N = 126) in Dutch Ipsos I&O Panel in 2023 (Van Es et al., 2025)
 - *Participants requested and downloaded their Data Download Packages (DDPs) from Netflix*
- PORT locally extracted relevant traces from the DDPs (Boeschoten et al., 2023)
 - *This includes viewing histories (i.e., titles of series episodes/movies viewed with timestamps)*
- In total, this resulted in 617,951 traces
 - *These contain 66,033 unique traces (i.e., different series episodes/movies)*
 - *Traces must be enriched as they do not contain any information on genres and runtimes*
- Development of a three-step workflow to link, enrich, and transform traces
 - *Pre-processing step (“extraction of title information”) is not considered in this presentation*
 - *Workflow is connected to The Movie Database (TMDB) API and automated in Python*
 - *It is evaluated by calculating quality indicators for each workflow step*

Workflow: (1) Linkage with TMDb Identifier I

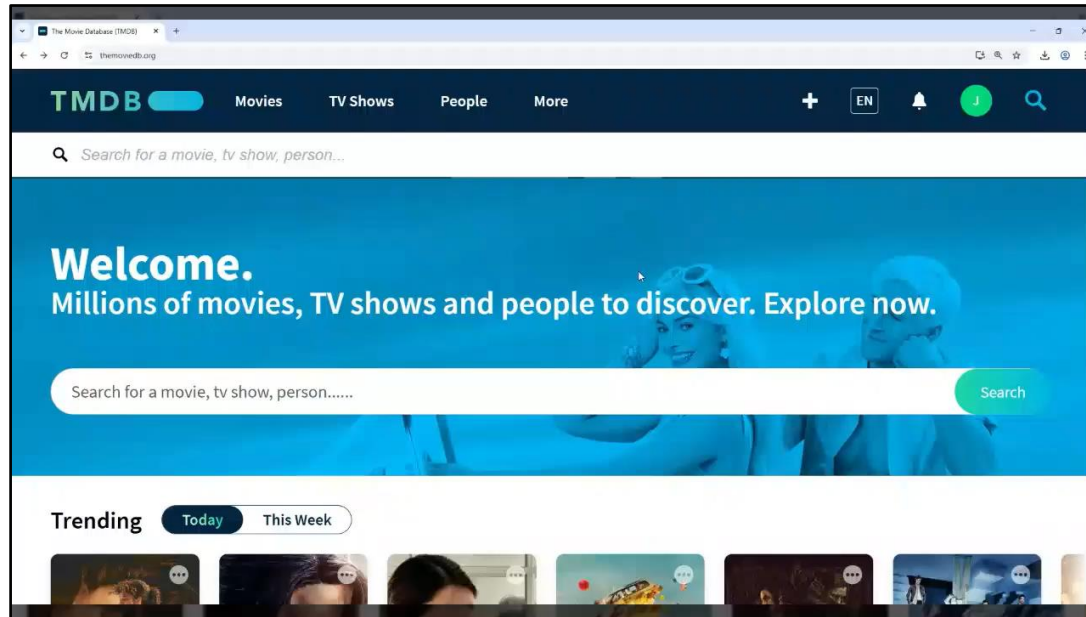
Series: "The Office", Season: 2, Title: "Motivation", Episode: 4



Which one is it?!

Workflow: (1) Linkage with TMDb Identifier II

Series: "The Office", Season: 2, Title: "Motivation", Episode: 4



TMDb ID: 2996

Workflow: (2) Enrichment with Auxiliary Information


TMDB-ID: 2996, Series: "The Office", Season: 2, Title: "Motivation", Episode: 4

Genre



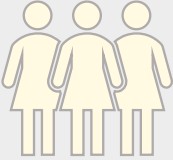
Comedy

Runtime




27 minutes

Cast



Ricky Gervais
(male), Martin
Freeman (male),
...

Country



United Kingdom

Workflow: (3) Transformation into Genre Measurements

	Duration	Frequency
Absolute measurement	Michael views <u>2 hours</u> of "comedy" content per week.	Michael views <u>4 episodes/movies</u> of "comedy" content per week.
Relative measurement	Michael spends <u>50 percent</u> of his Netflix <u>time</u> on "comedy" content.	Of all <u>episodes/movies</u> viewed by Michael, <u>65 percent</u> contain "comedy" content.
Threshold for counting a "view" was viewing at least <u>0%</u> of an episode's/movie's runtime (alternatively: <u>70%</u> or <u>95%</u>).		

Hi, I'm Michael. I like to laugh.



Results: (1) Linkage with TMDb Identifier

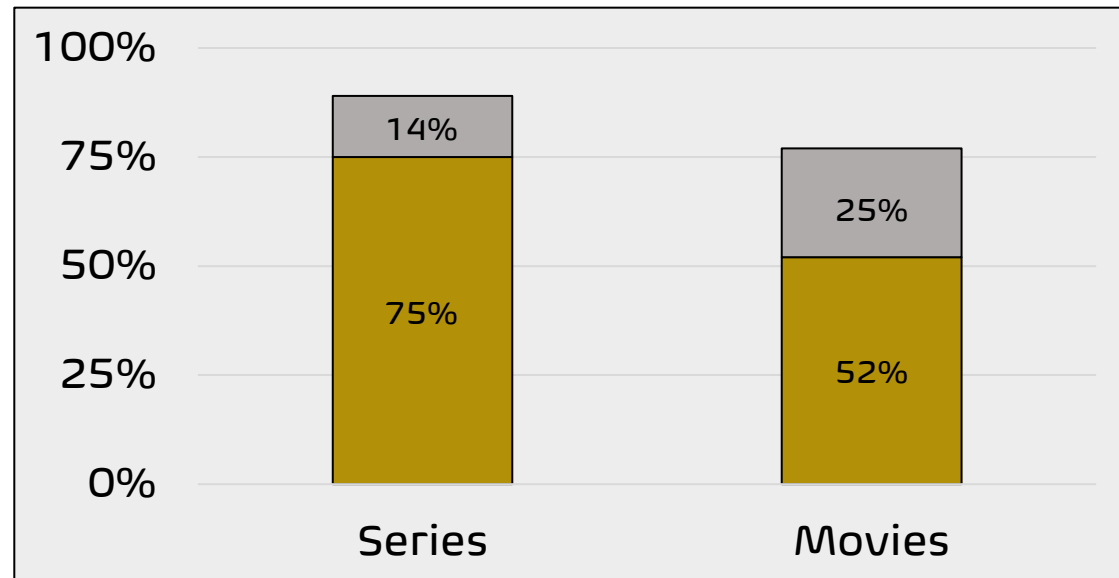
Quality indicator

Share of exact and probability-based matches, considering only unique traces

Calculation

$$\frac{\text{exact matches} + \text{prob matches}}{\text{exact matches} + \text{prob matches} + \text{non matches}}$$

Fig 1. Share of exact (gold) and probability-based (silver) matches (in %)



Results: (2) Enrichment with Auxiliary Information

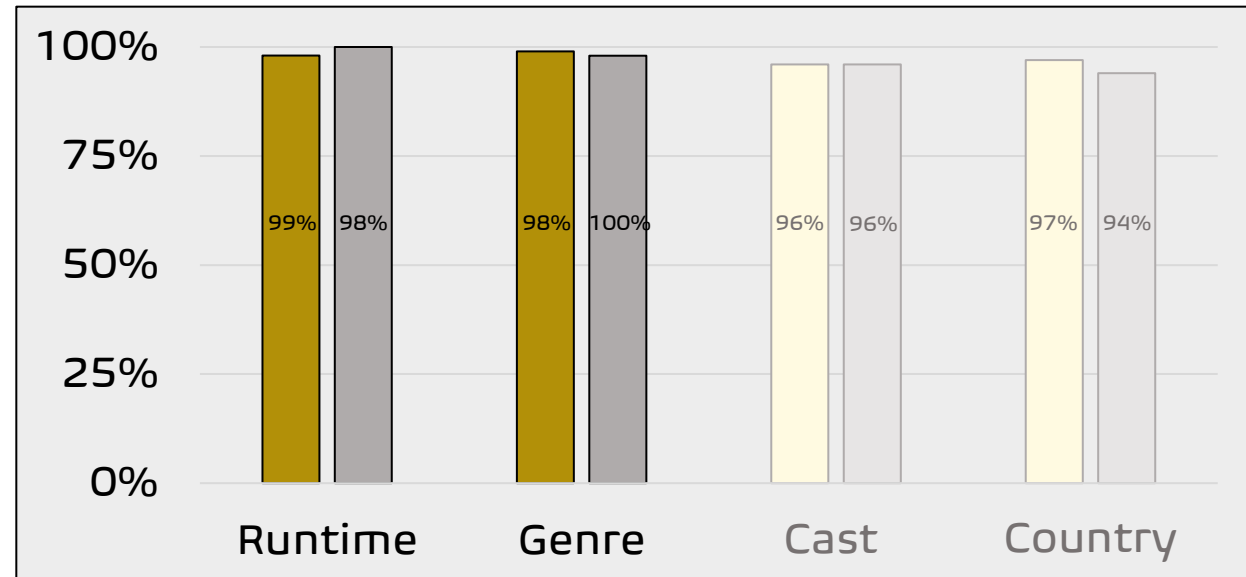
Quality indicator

Share of enriched traces,
considering only matched traces

Calculation

$$\frac{\text{enriched traces}}{\text{enriched traces} + \text{non enriched traces}}$$

Fig 2. Share of enriched series (gold) and movie traces (silver) (in %)



Results: (3) Transformation into Genre Measurements

Quality indicator

Robustness to variations
in analytical decisions:

0% vs. 70% vs. 95% threshold

Duration vs. frequency

Absolute vs. relative measurement

Calculation

$$r_{xy} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

Fig 3. Pairwise correlation coefficients between measurements of "comedy" genre

abs. dur. (0%)	1.00	1.00	0.98	0.89	0.89	0.89	0.60	0.60	0.59	0.60	0.59	0.58
abs. dur. (70%)	1.00	1.00	0.99	0.88	0.89	0.89	0.61	0.60	0.60	0.60	0.59	0.58
abs. dur. (95%)	0.98	0.99	1.00	0.83	0.85	0.87	0.60	0.59	0.60	0.58	0.57	0.57
abs. freq. (0%)	0.89	0.88	0.83	1.00	0.98	0.96	0.65	0.64	0.62	0.71	0.70	0.68
abs. freq. (70%)	0.89	0.89	0.85	0.98	1.00	0.99	0.65	0.65	0.64	0.71	0.70	0.69
abs. freq. (95%)	0.89	0.89	0.87	0.96	0.99	1.00	0.66	0.66	0.66	0.71	0.70	0.70
rel. dur. (0%)	0.60	0.61	0.60	0.65	0.65	0.66	1.00	0.99	0.97	0.95	0.94	0.92
rel. dur. (70%)	0.60	0.60	0.59	0.64	0.65	0.66	0.99	1.00	0.99	0.94	0.95	0.94
rel. dur. (95%)	0.59	0.60	0.60	0.62	0.64	0.66	0.97	0.99	1.00	0.92	0.93	0.95
rel. freq. (0%)	0.60	0.60	0.58	0.71	0.71	0.71	0.95	0.94	0.92	1.00	0.98	0.96
rel. freq. (70%)	0.59	0.59	0.57	0.70	0.70	0.70	0.94	0.95	0.93	0.98	1.00	0.98
rel. freq. (95%)	0.58	0.58	0.57	0.68	0.69	0.70	0.92	0.94	0.95	0.96	0.98	1.00
	abs. dur. (0%)	abs. dur. (70%)	abs. dur. (95%)	abs. freq. (0%)	abs. freq. (70%)	abs. freq. (95%)	rel. dur. (0%)	rel. dur. (70%)	rel. dur. (95%)	rel. freq. (0%)	rel. freq. (70%)	rel. freq. (95%)

Conclusion

- High volume of traces (n = 617,951) necessitates an automated workflow
- Linking traces to the TMDB identifier represents the biggest error source
 - *Between 10% and 25% of traces could not be linked at all (-> missing data error)*
 - *Probability-based linkages (15% to 25%) may result in erroneous genre attributions*
- Analytical decisions differ in their impact on research outcomes
 - *Using durations, frequencies, and different viewing thresholds results in similar measurements*
 - ***But:*** *Absolute measurements represent genre exposure (-> media effects research)*
 - ***In contrast:*** *Relative measurements represent genre preferences (-> research on media choices)*

Many thanks for your attention!

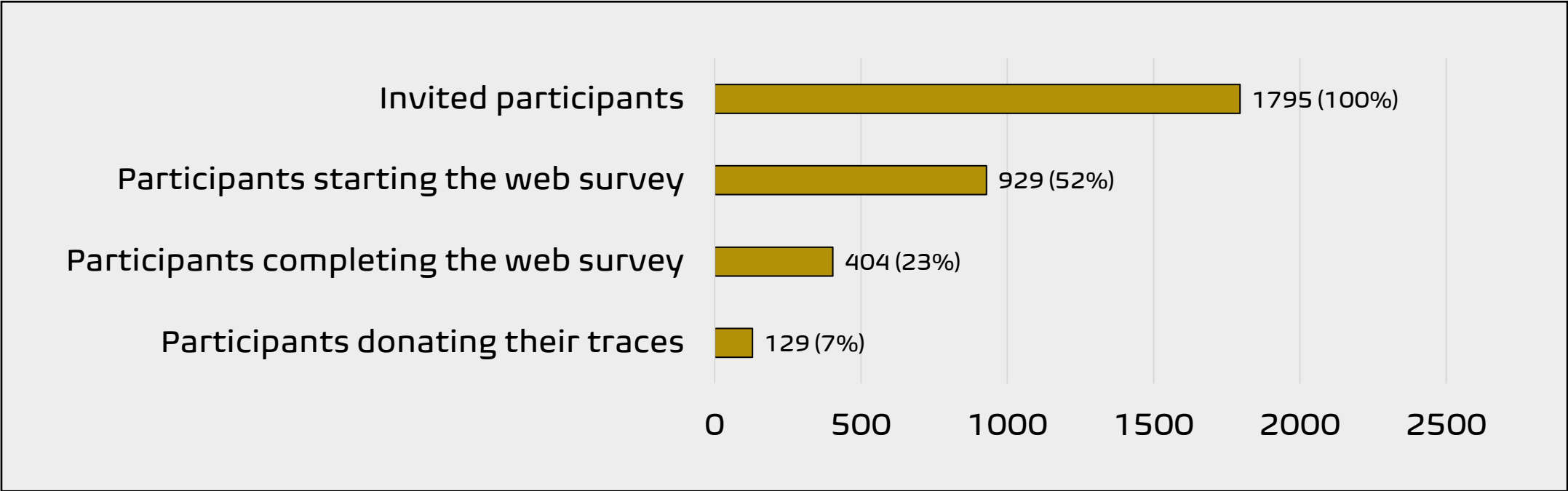
@jclaass.bsky.social
claassen@dzhw.eu

Literature

- Baghal, T. A., Belli, R. F., Phillips, A. L., & Ruther, N. (2014). What are you doing now? Activity-level responses and recall failures in the American Time Use Survey. *Journal of Survey Statistics and Methodology*, 2(4), 519-537. <https://doi.org/10.1093/jssam/smu020>
- Kaur, M. H., & Ashfaq, D. R. (2023). The impact of Netflix on viewer behaviour and media consumption: An exploration of the effects of streaming services on audience engagement and entertainment preferences. *Journal of Media, Culture and Communication*, 3(4), 9-23. <https://doi.org/10.55529/jmcc.34.9.23>
- Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, 72(5), 847-865. <https://doi.org/10.1093/poq/nfn063>
- Lobato, R., & van Es, K. (2025). Video-on-Demand Research: New Methods, Old Questions. *Media Industries*, 12(1). <https://doi.org/10.3998/mij.6359>
- Netflix (2024). 2023 annual report. https://s22.q4cdn.com/959853165/files/doc_financials/2023/ar/Netflix-10-K-01262024.pdf
- Ochoa, C., & Revilla, M. (2025). Variability of a job search indicator induced by operationalization decisions when using digital traces from a meter. *PLoS One*, 20(12), Article e0338894. <https://doi.org/10.1371/journal.pone.0338894>
- Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, 5, 1535-1547. <https://doi.org/10.1038/s41562-021-01117-5>
- Tana, J., Eirola, E., & Nylund, M. (2019). When is prime-time in streaming media platforms and video-on-demands services? New media consumption patterns and real-time economy. *European Journal of Communication*, 35(2), 108-125. <https://doi.org/10.1177/0267323119894482>
- Wedel, L., Ohme, J., & Araujo, T. (2024). Augmenting data download packages: Integrating data donations, video metadata, and the multimodal nature of audio-visual content. *Methods, Data, Analyses*, 32. <https://doi.org/10.12758/mda.2024.08>

Appendix I: Study Participation

Fig A1. Metrics on study participation



Appendix II: Sample Characteristics

Tab A2. Sociodemographic sample characteristics

	Web survey completes	Donation of traces
Age (mean)	49	42
Female gender (%)	48	39
<i>Education</i>		
Low education (%)	23	15
Medium education (%)	43	46
High education (%)	34	40
N	404	129

Appendix III: Workflow Summary

- (1) Extraction of title information (using regular expressions)
 - *Parsing the series or movie title, season number, episode title, and episode number*
- (2) Linkage with unique identifier in The Movie Database (TMDB)
 - *Movies are matched based on title; series are matched based on series and episode title*
 - *Probability-based matching if exact matching is not possible*
- (3) Enrichment with auxiliary information on genres and runtimes
 - *Retrieved through the TMDB API using the unique identifier*
 - *Information on cast and production country are not substantially analyzed in this presentation*
- (4) Transformation of traces into individual-level genre measurements
 - *Threshold for counting a “view”: 0% vs. 70% vs. 95% of runtime*
 - *Aggregating duration of views vs. frequency of views*
 - *Creating absolute vs. relative measurements (i.e., proportional to overall Netflix usage)*

Appendix IV: Extraction of Title Information

“The Office: Season 2: Motivation (Episode 4)”

Series Season Title Episode

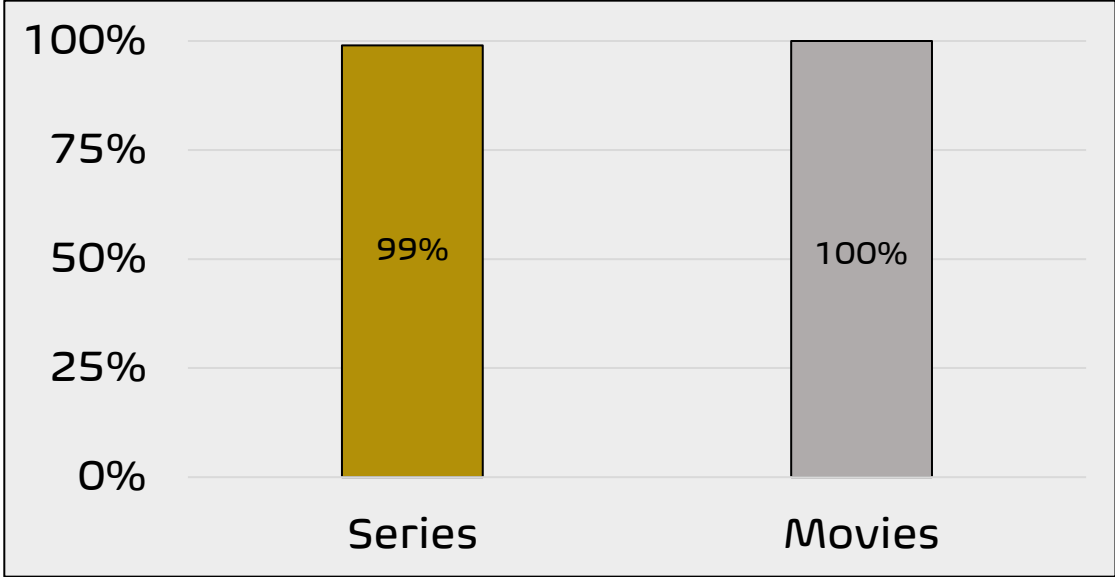
Appendix IV: Extraction of Title Information

Quality indicator
Share of parsed series and movie traces, considering only unique traces

Calculation

$$\frac{\text{parsed titles}}{\text{parsed titles} + \text{non parsed titles}}$$

Fig A4. Share of parsed series and movie titles (in %)



Appendix V: "Drama" Measurements

Quality indicator

Robustness to variations
in transformation rules:
0% vs. 70% vs. 95% threshold
Duration vs. frequency
Absolute vs. proportional

Calculation

$$r_{xy} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

Fig A5. Pairwise correlation coefficients between measurements of "drama" genre

abs. dur. (0%)	1.00	1.00	0.97	0.91	0.92	0.92	0.43	0.40	0.43	0.41	0.37	0.41
abs. dur. (70%)	1.00	1.00	0.97	0.90	0.92	0.93	0.43	0.40	0.42	0.41	0.37	0.41
abs. dur. (95%)	0.97	0.97	1.00	0.81	0.83	0.91	0.42	0.39	0.44	0.40	0.35	0.42
abs. freq. (0%)	0.91	0.90	0.81	1.00	0.97	0.92	0.46	0.43	0.45	0.47	0.41	0.45
abs. freq. (70%)	0.92	0.92	0.83	0.97	1.00	0.95	0.48	0.44	0.45	0.49	0.43	0.46
abs. freq. (95%)	0.92	0.93	0.91	0.92	0.95	1.00	0.50	0.47	0.52	0.51	0.45	0.53
rel. dur. (0%)	0.43	0.43	0.42	0.46	0.48	0.50	1.00	0.98	0.93	0.96	0.94	0.90
rel. dur. (70%)	0.40	0.40	0.39	0.43	0.44	0.47	0.98	1.00	0.94	0.94	0.96	0.92
rel. dur. (95%)	0.43	0.42	0.44	0.45	0.45	0.52	0.93	0.94	1.00	0.87	0.87	0.96
rel. freq. (0%)	0.41	0.41	0.40	0.47	0.49	0.51	0.96	0.94	0.87	1.00	0.97	0.90
rel. freq. (70%)	0.37	0.37	0.35	0.41	0.43	0.45	0.94	0.96	0.87	0.97	1.00	0.91
rel. freq. (95%)	0.41	0.41	0.42	0.45	0.46	0.53	0.90	0.92	0.96	0.90	0.91	1.00
abs. dur. (0%)												
abs. dur. (70%)												
abs. dur. (95%)												
abs. freq. (0%)												
abs. freq. (70%)												
abs. freq. (95%)												
rel. dur. (0%)												
rel. dur. (70%)												
rel. dur. (95%)												
rel. freq. (0%)												
rel. freq. (70%)												
rel. freq. (95%)												

Appendix VI: "Crime" Measurements

Quality indicator

Robustness to variations
in transformation rules:
0% vs. 70% vs. 95% threshold
Duration vs. frequency
Absolute vs. proportional

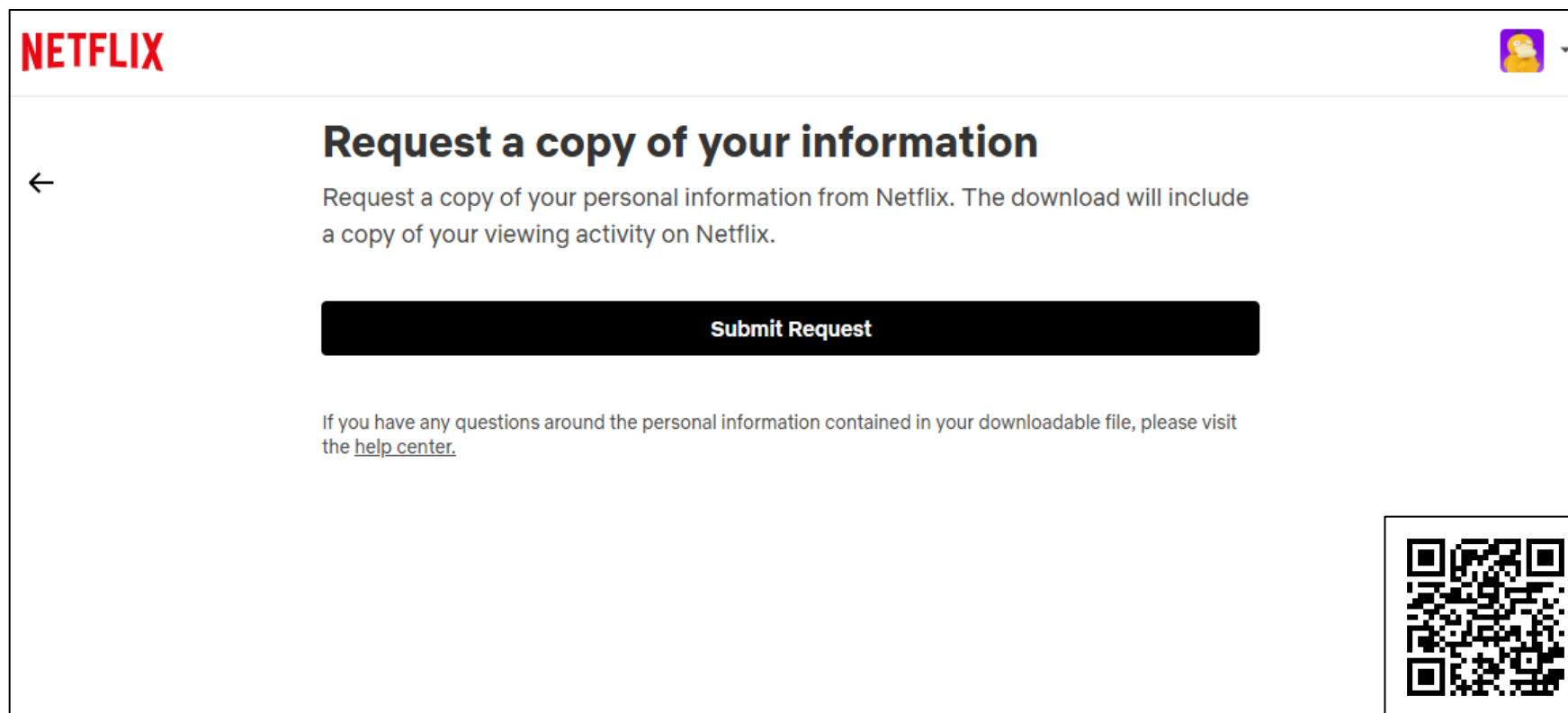
Calculation

$$r_{xy} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

Fig A6. Pairwise correlation coefficients between measurements of "crime" genre

abs. dur. (0%)	1.00	1.00	0.95	0.95	0.95	0.93	0.47	0.46	0.41	0.44	0.43	0.39
abs. dur. (70%)	1.00	1.00	0.96	0.93	0.95	0.94	0.48	0.48	0.43	0.45	0.44	0.40
abs. dur. (95%)	0.95	0.96	1.00	0.85	0.88	0.94	0.50	0.49	0.46	0.46	0.45	0.43
abs. freq. (0%)	0.95	0.93	0.85	1.00	0.98	0.92	0.51	0.49	0.44	0.49	0.47	0.42
abs. freq. (70%)	0.95	0.95	0.88	0.98	1.00	0.96	0.54	0.52	0.48	0.52	0.50	0.46
abs. freq. (95%)	0.93	0.94	0.94	0.92	0.96	1.00	0.59	0.57	0.55	0.57	0.55	0.53
rel. dur. (0%)	0.47	0.48	0.50	0.51	0.54	0.59	1.00	0.99	0.98	0.98	0.97	0.96
rel. dur. (70%)	0.46	0.48	0.49	0.49	0.52	0.57	0.99	1.00	0.98	0.97	0.98	0.96
rel. dur. (95%)	0.41	0.43	0.46	0.44	0.48	0.55	0.98	0.98	1.00	0.96	0.96	0.98
rel. freq. (0%)	0.44	0.45	0.46	0.49	0.52	0.57	0.98	0.97	0.96	1.00	0.99	0.98
rel. freq. (70%)	0.43	0.44	0.45	0.47	0.50	0.55	0.97	0.98	0.96	0.99	1.00	0.98
rel. freq. (95%)	0.39	0.40	0.43	0.42	0.46	0.53	0.96	0.96	0.98	0.98	0.98	1.00
abs. dur. (0%)							rel. dur. (0%)					
abs. dur. (70%)							rel. dur. (70%)					
abs. dur. (95%)							rel. dur. (95%)					
abs. freq. (0%)							rel. freq. (0%)					
abs. freq. (70%)							rel. freq. (70%)					
abs. freq. (95%)							rel. freq. (95%)					

Appendix VII: Digital Data Donation



The screenshot shows the Netflix user interface for requesting personal information. At the top left is the 'NETFLIX' logo in red. At the top right is a user profile icon. The main heading is 'Request a copy of your information' in bold black text. Below the heading is a left-pointing arrow and a paragraph: 'Request a copy of your personal information from Netflix. The download will include a copy of your viewing activity on Netflix.' A large black button with the text 'Submit Request' is centered below the paragraph. Further down, there is another paragraph: 'If you have any questions around the personal information contained in your downloadable file, please visit the [help center](#).' In the bottom right corner, there is a QR code.