Inverse Chilling Effect: General Data Protection Regulation application and Wikipedia Use

Gonxhe Idrizi gonxhe.idrizi@epfl.ch
Matthias Zeller matthias.zeller@epfl.ch
Nicolas D'Argenlieu nicolas.dargenlieu@epfl.ch

Abstract

This paper is an extension research to the article *Chilling effects: online surveillance and Wikipedia use* from Jonathan W. Penney. It analyses the traffic of german Wikipedia articles related to the General Data Protection Regulation (GDPR) around the announcement of the regulation in April 2016 and the day it became effective in May 2018. A traffic increase after the regulation announcements would give ground to the hypothesis of a reverse chilling effect on the European population. We attempt to provide answers to this research question through the pageviews trends exploration before and after the aforementioned dates using two Interrupted Time Series (ITS). The obtained results are unsatisfactory evidence to the reverse chilling effect hypothesis following the announcement and adoption of the GDPR regulation.

1 Introduction

On the 27th of April 2016, the European Parliament adopted the General Data Protection Regulation (GDPR)[1], which became effective the 25th May 2018. This european regulation sets principles applicable to any business or institution on the management of personal data inside the European Union and European Economic Area. More precisely it ensures data safeguard and privacy, and provides the user with more control and information on its data. We explore the hypothesis of a reverse chilling effect (i.e. growth of interest) in Europe following the adoption of the regulation and its start of effectiveness. To answer this, we build a Wikipedia dataset with german Wikipedia articles related to GDPR. We focus on the German domain only as its vast majority of speakers reside in Europe. Therefore we assume this traffic best represents a European reaction to this event. We use article pageviews as a metric to explore the pre-and post-trends of the aforementioned dates. We first compare the means of monthly aggregated view counts. Then using an Interrupted Time Series (ITS) design, the trends of the GDPR-related articles are illustrated through segmented linear regressions with 95% confidence intervals. To further strenghten our results, we compare this treatment group to a control group consisting of the most popular articles for the period.

This study is inspired by the article *Chilling effects: online surveillance and Wikipedia use* (Penney, 2016), demonstrating the feasible usage of Wikipedia to support the evidence of a chilling effect following Edward Snowden's revelation in June 2013 (Lyon, 2014).

2 Data

2.1 Data Scraping

We choose to use the Wikimedia REST API[2] via the package mwviews[3] to retrieve Wikipedia pageviews. The API provides statistics on the daily or monthly view counts of a chosen article for a given time range. However, the oldest data we were able to retrieve with this method dates to the 1st July 2015, whereas our selected period of interest starts in April 2015. An alternative was to use raw dump files from the pagecounts dataset[4] containing the hourly views of articles in all Wikipedia domains. However we noted that this dataset is no longer maintained.

Considering this, we needed to check data consistency between both sources. We down-

[1] https://gdpr-info.eu/

[2] https://wikimedia.org/api/rest_v1/

[3] https://github.com/mediawiki-utilities/python-mwviews

[4] https://wikitech.wikimedia.org/wiki/Analytics/Archive/Data/Pagecounts-raw

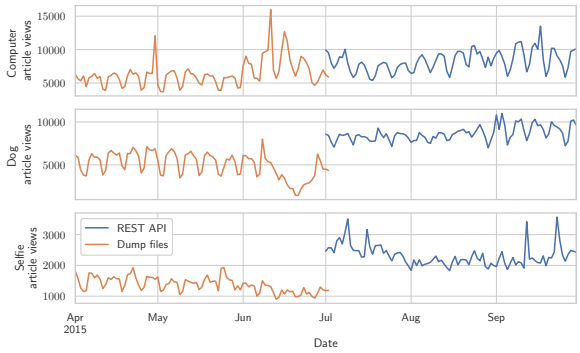

Figure 1: Sequential comparison of (deprecated) Wikipedia pagecounts and Wikipedia REST API pageviews for three selected articles of the *en* Wikipedia project: Computer, Dog and Selfie. Data has daily granularity.

loaded mainstream articles `Dog`, `Computer` and `Selfie` from 1st April 2015 to 2nd July 2015 as depicted in Figure 1 to join to our datasets. Due to the important trend discrepancies, we conclude that one or both sources are not reliable. We choose Wikipedia's API for maintainability reasons.

2.2 Keywords lists

Extraction of keywords

We selected GDPR-related Wikipedia articles by gathering english keywords found in newspapers and media describing GDPR regulation, and by selecting the corresponding articles from english Wikipedia category names[5]. Articles were cross-validated among the group and translated in german to be manually paired with existing german articles.

Pre-processing of keywords

The daily pageviews visualization, in Figure 4 of the Appendix, features a yearly drop in the end of December for most articles. This month thus constitutes an outlier in the computation of monthly trends for aggregated pageviews. Therefore we remove it for all years in both treatment and control groups. There are many plausible interpretations for this phenomena, one being that a majority of visitors are employees of security and IT departments who are in holidays on December.

In the treatment group, we removed the articles *Individualrecht, Datenvernichtung, Zensur im Internet, Vorratsdatenspeicherung* as these display a unique unexplained random spike in views, negatively impacting our aggregation of all articles. *Cyber-terrorismus* was discarded as well, as it is sensitive to cyber attacks events, shows a very noisy pageview count, and its relatedness to GDPR is limited.

In the control group, we removed articles featuring a single important spike, assuming those represent an unrelated important event (Table 6). We also removed TV show articles as their trend strongly depends on the show's releases.

[5] `https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories`

3 Methods and analysis

Treatment and control means and ITS datasets were created over a 22 months period (2 years excluding Decembers), of monthly aggregated articles pageviews.

3.1 Means comparison

We measure the mean number of views for the control and treatment groups before and after the interruption event. If there is a lasting effect following the event, then the total views' average of the treatment group should depict a variation.

3.2 Segmented regression

For the second analysis, we computed a segmented regression for the Interrupted Time Series (ITS) of the treatment and control datasets. This framework compares the pre- and post-intervention event trends.

This method comes from health economy analyses such as in (Lagarde and Palmer, 2008). They propose the following equation :

$$Y_t = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot \mathrm{I}_{\{t>t_0\}} + \beta_3 \cdot \mathrm{I}_{\{t>t_0\}}(t - t_0) + \varepsilon_t,$$

with $\mathrm{I}_{\{\cdot\}}$ the indicator variable, t_0 the month of interruption event. β_2 and β_3 are often referred to as intervention and postslope respectively.

4 Results

4.1 Means comparison

For the law adoption event in April 2016, the mean pageviews of the GDPR-related articles increases by 1.26% when considering a 12-months time window, but decreases by 12.0% for a 6-months time window. The overlapping of the error bars, as shown in Figure 2, do not suggest a significant variation of views means following the

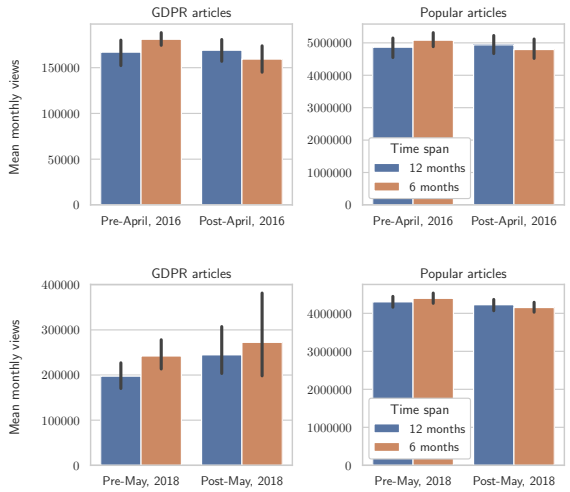

Figure 2: Barplots of the mean page views over two different time windows (6 and 12 months) around the GDPR adoption date (top panel) and law enforcement date (bottom panel)

law adoption. The control group's means behave similarly.

For the regulation enforcement in May 2018, (Figure 2) GDPR articles show a 24.0% and 12.5% increase in means for the long and short time windows respectively. Note the size of the confidence intervals which indicates a strong uncertainty in the distribution of means, especially in the six months following May 2018.

4.2 Segmented regression

For April 2016, the p-values of the F-test statistic for both GDPR-related articles (0.128, Table 1 in the Appendix) and popular articles (0.351) suggest that this ITS model cannot be considered as statistically different from an intercept-only model (i.e. the mean). This is further confirmed by the poor model fit ($R^2_{\text{adj}} = 0.029$).

For May 2018, the model suggests an increasing pre-May trend ($\hat{\beta}_1 = +13,290$, p-value 3.7%), whereas the post-trend ($\beta_1 + \beta_3$) ranges from -47,615 to +12,300 under 95% confidence level (see Figure 3). The immediate 86,680 increase in pageviews is not statistically significant (p-value 0.12). Although the overall model is statistically significant (F-test p-value 0.0062), the 12th month stands out (May 2018), cumulating 523,996 views (2.42 times the mean number of views for the latter months). The control group does not exhibit this irregularity for month 12. Furthermore the post-slope trend cannot be inferred due to the large

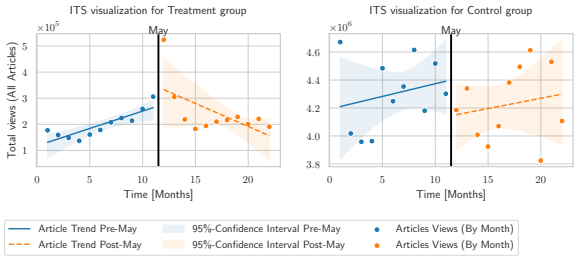

Figure 3: Segmented regression plots with GDPR law enforcement as interruption event. Left panel: GDPR-related articles. Right panel: control group.

confidence intervals.

5 Discussion

We establish that the ITS model is not suited for neither analysis of the GDPR events. There are very few months around the intervention event which have their total pageviews impacted by the interruption event (apart from the interruption month itself). The ITS design seems here limited. Indeed european laws are announced and the regulation cannot be affiliated with a sudden treatment. The pageviews trends could therefore have been impacted by prior information and one cannot identify an interruption as the source of long-lasting effect for GDPR. Furthermore one limitation has been the scattered monthly views for popular Wikipedia articles. It would have been relevant to deeper investigate for outliers' removal and data expansion in the control group.

6 Further research

Further research could adress the same question by focusing on other European communities and compare their view trends to the german community. This will allow to determine which countries were most concerned by the GDPR regulation.

7 Conclusion

An inverse chilling effect could not be assessed through our analysis, neither for GDPR regulation adoption nor its enforcement. However an immediate and abrupt spike in pageviews for GDPR-related articles is visible around May 2018.

References

Lagarde, M. and Palmer, N. (2008). The impact of user fees on health service utilization in low- and middle-income countries: how strong is the evidence? *Bulletin of the World Health Organization*, 86(11):839–848.

Lyon, D. (2014). Surveillance, Snowden, and Big Data: Capacities, consequences, critique. *Big Data & Society*, 1(2):2053951714541861. Publisher: SAGE Publications Ltd.

Penney, J. W. (2016). Chilling Effects: Online Surveillance and Wikipedia Use. *Berkeley Technology Law Journal*, 31(1):117. Number: IR.

Appendices

	coef	std err	t	P > |t|	[0.025	0.975]
Intercept	136800.0000	15100.000	9.077	0.000	105000.000	169000.000
time	6705.5714	2983.553	2.248	0.040	346.278	13100.000
intervention	-35140.0000	17700.000	-1.989	0.065	-72800.000	2517.504
postslope	-4409.5987	3507.190	-1.257	0.228	-11900.000	3065.799

Table 1: Segmented regression summary table of GDPR-related articles around date of GDPR adoption (April 2016). Adjusted R^2: 0.169, F-test p-value: 12.8%

	coef	std err	t	P > |t|	[0.025	0.975]
Intercept	4296000.0	353000.0	12.177	0.000	3540000.0	5050000.0
time	126300.0	69900.0	1.807	0.091	-22700.0	275000.0
intervention	-468400.0	414000.0	-1.132	0.275	-1350000.0	413000.0
postslope	-109600.0	82100.0	-1.335	0.202	-285000.0	65400.0

Table 2: Segmented regression summary table of popular articles around date of GDPR adoption (April 2016). Adjusted R^2: 0.029, F-test p-value: 35.1%.

	coef	std err	t	P > |t|	[0.025	0.975]
Intercept	117700.0	40000.000	2.940	0.009	33600.000	202000.0
time	13290.0	5903.536	2.251	0.037	885.086	25700.0
intervention	86680.0	53100.000	1.631	0.120	-24900.000	198000.0
postslope	-30930.0	8348.861	-3.705	0.002	-48500.000	-13400.0

Table 3: Segmented regression summary table of GDPR-related articles around date of GDPR enforcement (May 2018). Adjusted R^2: 0.403, F-test p-value: 0.618%.

	coef	std err	t	P > |t|	[0.025	0.975]
Intercept	4.192000e+06	173000.0	24.284	0.000	3830000.0	4550000.0
time	1.815000e+04	25500.0	0.713	0.485	-35300.0	71600.0
intervention	-2.544000e+05	229000.0	-1.111	0.281	-736000.0	227000.0
postslope	-3.462245e+03	36000.0	-0.096	0.924	-79100.0	72200.0

Table 4: Segmented regression summary table of popular articles around date of GDPR enforcement (May 2018). Adjusted R^2: -0.089, F-test p-value: 73.6%.

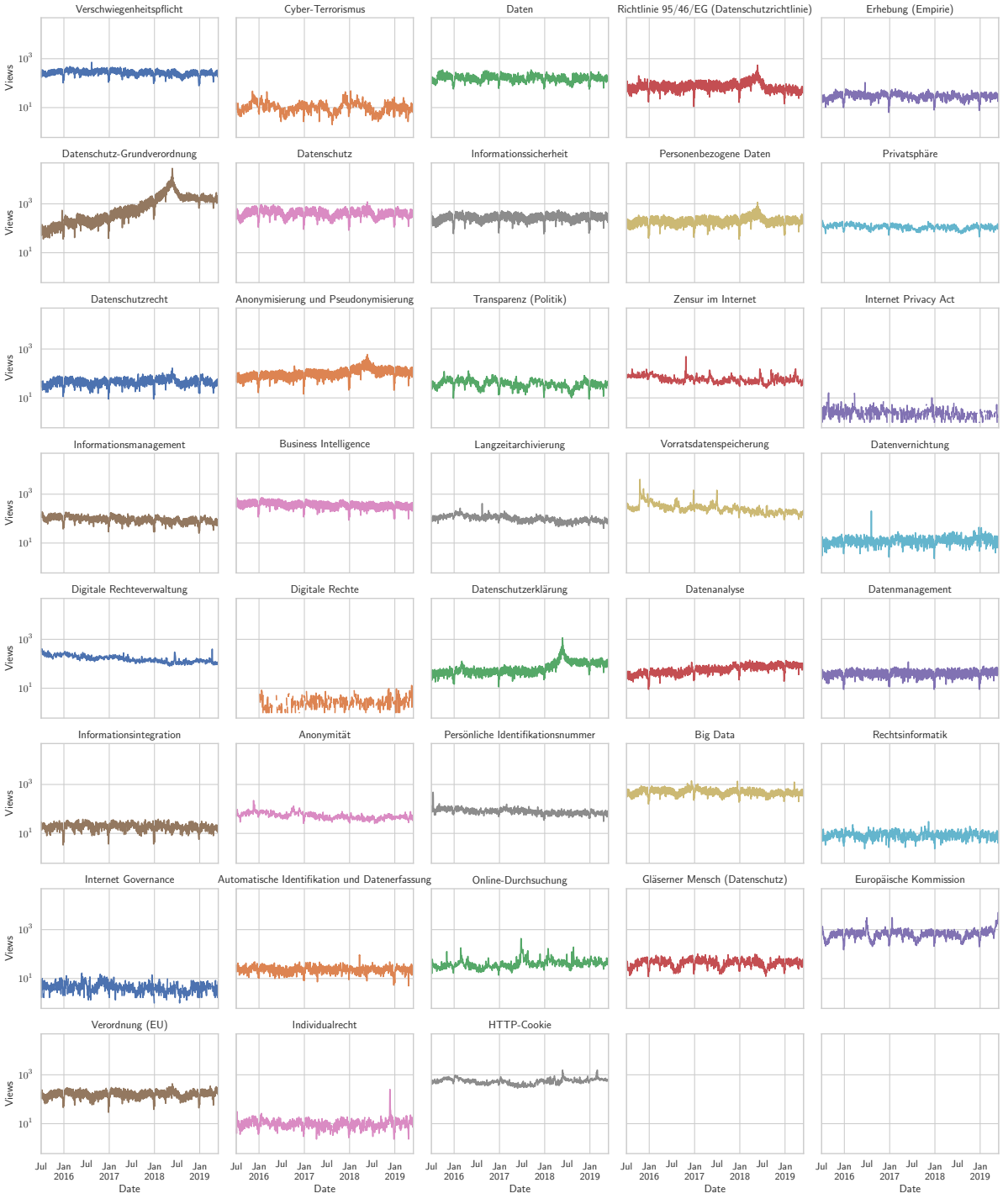

Figure 4: Daily views of GDPR-related articles. Data are smoothed with a three-day rolling average for clarity.

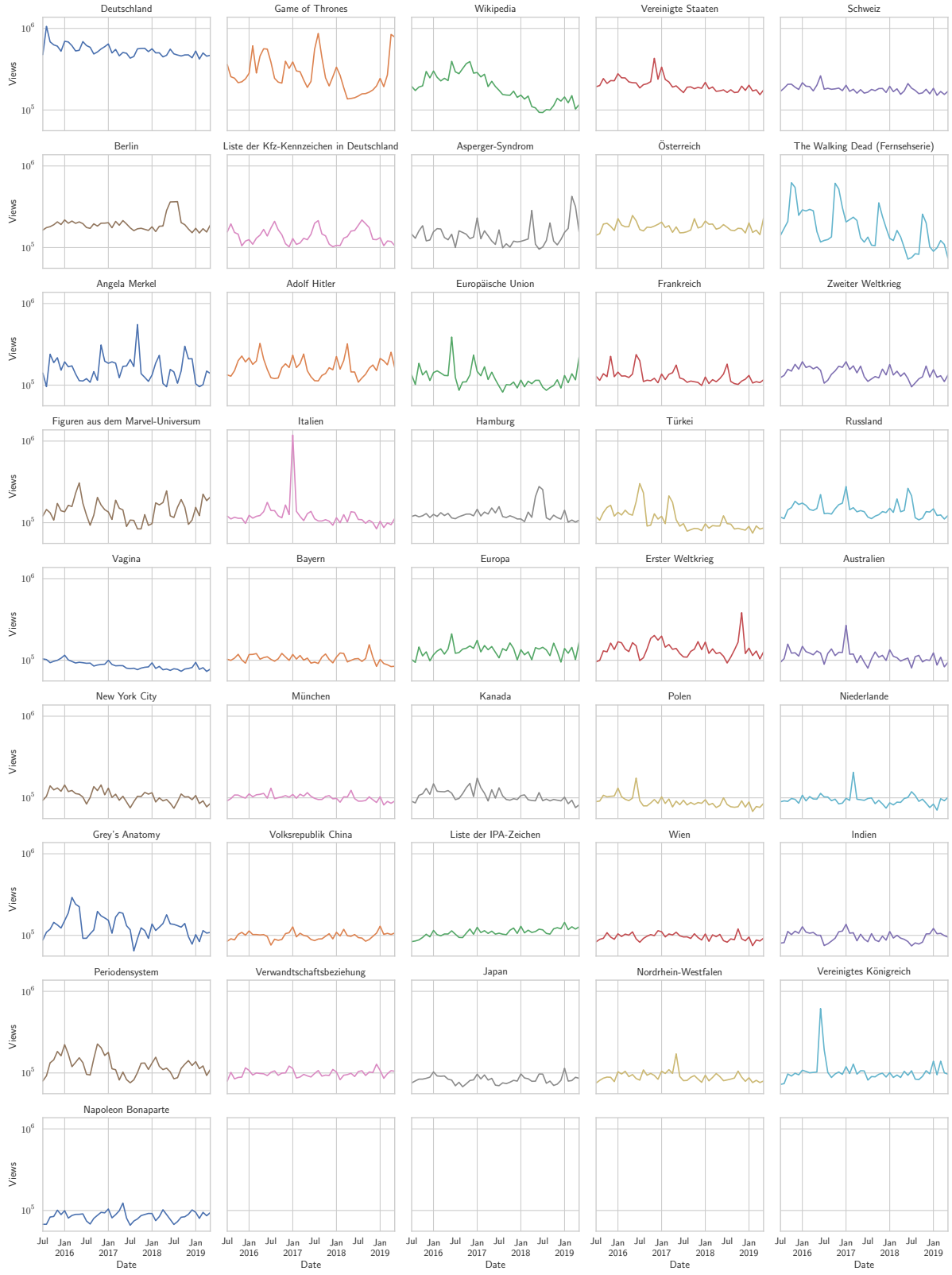

Figure 5: Monthly views of popular articles.

German Wikipedia article	Corresponding English article
Verschwiegenheitspflicht	Confidentiality
Daten	Data
Richtlinie 95/46/EG (Datenschutzrichtlinie)	Data Protection Directive
Erhebung (Empirie)	Data collection
Datenschutz-Grundverordnung	General Data Protection Regulation
Datenschutz	Information privacy
Informationssicherheit	Information security
Personenbezogene Daten	Personal data
Privatsphäre	Privacy
Datenschutzrecht	Privacy law
Anonymisierung und Pseudonymisierung	Pseudonymization
Transparenz (Politik)	Transparency (behavior)
Internet Privacy Act	Internet Privacy Act
Informationsmanagement	Information management
Business Intelligence	Business intelligence
Langzeitarchivierung	Digital preservation
Digitale Rechteverwaltung	Digital rights management
Digitale Rechte	Digital rights
Datenschutzerklärung	Privacy policy
Datenanalyse	Data analysis
Datenmanagement	Data management
Informationsintegration	Information integration
Anonymität	Anonymity
Persönliche Identifikationsnummer	Personal identification number
Big Data	Big data
Rechtsinformatik	Legal informatics
Internet Governance	Internet governance
Automatische Identifikation und Datenerfassung	Automatic identification and data capture
Online-Durchsuchung	Computer and network surveillance
Gläserner Mensch (Datenschutz)	Mass surveillance
Europäische Kommission	European Commission
Verordnung (EU)	Regulation (European Union)
HTTP-Cookie	HTTP cookie

Table 5: List of selected GDPR-related Wikipedia articles for the German Wikipedia (`de.wikipedia.org`) with their corresponding article in `en.wikipedia.org`.

German Wikipedia article	Corresponding English article
Zweiter Weltkrieg	World War II
Liste der Kfz-Kennzeichen in Deutschland	NaN
*Vereinigtes Königreich	United Kingdom
*Hamburg	Hamburg
Kanada	Canada
*Grey's Anatomy	Grey's Anatomy
Europa	Europe
Wien	Vienna
Schweiz	Switzerland
Europäische Union	European Union
Nordrhein-Westfalen	North Rhine-Westphalia
Angela Merkel	Angela Merkel
Periodensystem	Periodic table
Japan	Japan
Australien	Australia
Liste der IPA-Zeichen	Naming conventions of the International Phonet...
Bayern	Bavaria
Vagina	Vagina
Verwandtschaftsbeziehung	NaN
Russland	Russia
*Polen	Poland
Deutschland	Germany
*Italien	Italy
Napoleon Bonaparte	Napoleon
Volksrepublik China	China
Adolf Hitler	Adolf Hitler
Indien	India
Erster Weltkrieg	World War I
Asperger-Syndrom	Asperger syndrome
Figuren aus dem Marvel-Universum	Lists of Marvel Comics characters
München	Munich
Frankreich	France
Österreich	Austria
Wikipedia	Wikipedia
*Game of Thrones	Game of Thrones
Vereinigte Staaten	United States
*Niederlande	Netherlands
Türkei	Turkey
*The Walking Dead (Fernsehserie)	The Walking Dead (TV series)
*Berlin	Berlin
New York City	New York City

Table 6: List of popular Wikipedia articles for the German Wikipedia (`de.wikipedia.org`) with their corresponding article in `en.wikipedia.org`, if any. Articles marked with an asterisk (*) were excluded from the control group.