2. Zwischenbericht Datenspende

Personalisierung auf Googles Nachrichtenportal während der Bundestagswahl 2017

Autoren: Tobias D. Krafft, Michael Gamer, Katharina A. Zweig

Unser erster Zwischenbericht hat gezeigt, dass unsere Datenspenderinnen und -spender auf Googles allgemeiner Suchmaschine im Durchschnitt sehr viele Links gemeinsam in ihren Suchergebnissen sehen. Wir kamen daher zu dem Schluss, dass für eine algorithmisch erzeugte oder vertiefende „Filterblase“ nach Eli Parisers Theorie nicht viel Raum sei. Die Theorie von Eli Pariser fußt darauf, dass Algorithmen verschiedenen Nutzern auf dieselbe Suchanfrage personalisierte Suchergebnisse zuspielen, die inhaltlich dem entsprechen, was die Nutzer schon in früheren Suchergebnissen bevorzugt haben. Damit wäre es insbesondere bei politischen Themen möglich, dass Nutzerinnen und Nutzer nur noch eine eingeschränkte Sicht auf die Meinungsvielfalt zu einem Thema haben, die sich von dem, was andere sehen, unterscheidet.

In diesem Zwischenbericht geht es um eine analoge Untersuchung der Ergebnisse von Googles Nachrichtenportal, Google News.

Ergebnisse:

1. Auch auf dem Nachrichtenportal von Google ist der Raum für Personalisierung nicht groß, 4-5 der 20 Ergebnisse unterscheiden sich im Durchschnitt.
    a. Suchergebnisse für Personen haben im Durchschnitt 15,89 gleiche Ergebnisse, der Raum für eine mögliche Personalisierung liegt also bei gut 4 Einträgen.
    b. Die Parteien weisen auch hier einen etwas niedrigeren durchschnittlichen Mittelwert von 14.33 auf, woraus sich ein Raum für eine mögliche Personalisierung von 5-6 Einträgen ergibt.
2. Die Abweichung zwischen eingeloggten und ausgeloggten Nutzern liegt bei den Parteien im Durchschnitt bei unter 0,2 Ergebnissen. Bei den Suchen nach Personen erhalten eingeloggte Nutzer jedoch im Durchschnitt 0,85 mehr gleiche Ergebnisse, die ausgelieferten Ergebnisse sind sich also ähnlicher.
3. Bei Suchen auf der regulären Google-Suchmaschine sind viele URLs dauerhaft in den Suchergebnissen vertreten, wie z.B. der jeweilige Wikipedia-Eintrag oder die persönlichen Webseiten. Beim Nachrichtenportal von Google ist die Situation wesentlich dynamischer: Eine URL wird im Durchschnitt weniger als 2 Tage an Nutzer ausgerollt. Jedoch erreichen vereinzelte ULRs deutlich längere Anzeigezeiträume.

1 Einleitung

Gerade im Zeitalter der Digitalisierung, in dem es immer leichter und günstiger geworden ist, Nachrichten zu verfassen und zu verteilen (Flaxman et al., 2016), muss es der Gesellschaft ein wichtiges Anliegen sein, sich mit den Gefahren durch algorithmisch erzeugte und vertiefte Filterblasen zu beschäftigen. Als Filterblasen werden im Allgemeinen Umgebungen bezeichnet, die den darin gefangenen Personen nur eine stark selektierte Weltsicht erlauben. Eine solche Filterblase kann es prinzipiell auch im Analogen geben, wenn Menschen ihre Nachrichten z.B. von nur einer Tageszeitung erhalten, die zudem politisch stark gefärbt ist. Im Jahre 2011 äußerte aber Eli Pariser in seinem Buch „Filter Bubble – Wie wir im Internet entmündigt werden“ (Hanser Verlag, München) den Verdacht, dass *personalisierte Algorithmen*, die über die Reihenfolge von angezeigten Links und Nachrichten entscheiden, im Bereich von Suchmaschinen und sozialen Netzwerken solche Filterblasen vertiefen könnten oder sogar erzeugen könnten. Personalisierende Algorithmen lernen vom bisherigen Nutzerverhalten und kategorisieren Nutzer daraufhin. Bei weiterer Nutzung verändert die Kategorisierung des Nutzers oder der Nutzerin die Reihenfolge, in der die Links oder Nachrichten angezeigt werden, so dass die vermeintlich für den Nutzer interessantesten Links oder Nachrichten nach vorne gerückt werden. Parisers Argumentation beruhte nun (vereinfachend dargestellt) darauf, dass durch ein wiederholtes Anklicken von insbesondere politischen Nachrichten aus einem bestimmten Spektrum, der Algorithmus immer wieder Nachrichten aus demselben Spektrum bevorzugt. Ein solches Verhalten könnte große Auswirkungen auf die Meinungsbildung der Individuen und die Diskussionsfähigkeit innerhalb einer Gesellschaft haben. Es ist hier wichtig festzuhalten, dass die Theorie von Pariser ganz wesentlich darauf fußt, dass es eine Personalisierung der Suchergebnisreihenfolgen (oder Nachrichtenreihenfolgen auf sozialen Netzwerken) gibt. Wenn also sehr viele Nutzer fast immer dieselben Ergebnisse bekommen, bleibt nicht viel Raum für Personalisierung und damit keine Grundlage für eine algorithmisch erzeugte Filterblase. Ohne große Varianz der Suchergebnisse also keine Basis für eine Filterblase. Hier setzt unser Projekt an.

Das von sechs Landesmedienanstalten geförderte Projekt #Datenspende: Google und die Bundestagswahl 2017hat auf Google in den Wochen vor der Bundestagswahl die Suchergebnisse der ersten Seite zu politisch relevanten Suchanfragen gesammelt. Dazu haben die Kooperationspartner von AlgorithmWatch und das Algorithm Accountability Lab am Fachbereich Informatik an der TU Kaiserslautern mit Hilfe eines von der Firma *lokaler* gebauten Plugins alle 4 Stunden 16 vorher festgelegte Suchbegriffe automatisch auf Googles Suchmaschine eingeben lassen. Die 16 Suchbegriffe finden sich in Tabelle 1 und Tabelle 2.

SUCHBEGRIFFE PERSONEN
Dietmar Bartsch
Alexander Gauland
Katrin Göring-Eckardt
Christian Linder
Angela Merkel
Cem Özdemir
Martin Schulz
Sahra Wagenknecht
Alice Weidel

Tabelle 1: Suchbegriffe mit Namen von Politikern und Politikerinnen

SUCHBEGRIFFE PARTEIEN
AFD
BÜNDNIS90/DIE GRÜNEN
CDU
CSU
DIE LINKE
FDP
SPD

Tabelle 2: Suchbegriffe mit Namen von Parteien

Über den Verlauf des Projektes haben sich fast 4400 Personen das Plugin für die Browser Chrome oder Firefox heruntergeladen – zwischen 300 bis 600 von diesen Personen waren zu den Suchzeitpunkten um 12 Uhr, 16 Uhr und 20 Uhr aktiv und spendeten uns die jeweils für sie von Google berechneten Ergebnisse der ersten Suchergebnisseite. Die Suchanfragen wurden jeweils auf der normalen Google Suchmaschine und auf Googles Nachrichtenportal gestellt. Basierend auf diesen Daten können wir erstmalig Aussagen über die Diversität der ausgerollten Nachrichtenseiten sowie der Homogenität für die einzelnen Nutzer treffen.

Disclaimer: Da die Nutzerinnen und Nutzer nicht Teil eines repräsentativen Samples sind, sondern durch Berichterstattungen motiviert wurden, ist davon auszugehen, dass die Nutzergruppe nicht repräsentativ ist. Sie könnte z.B. bezüglich des Alters oder Bildungsstands homogener sein als die Gruppe aller deutschen Internetnutzer. Die nachfolgenden Resultate sind daher nicht abschließend, aber in ihrer Form so eindeutig, dass wir zuversichtlich sind, dass sich die Resultate im Großen und Ganzen auch bei einem repräsentativ ausgewählten Sample wiederholen würden.

In diesem zweiten Zwischenbericht werden nun die Ergebnisse der Suchanfragen auf dem Nachrichtenportal von Google (www.news.google.de) ausgewertet. Im Unterschied zur gewöhnlichen Google Suchanfrage mit einer Anzeige von 9 bis 10 Treffern für jede Suchanfrage erhält der Nutzer hier regelmäßig 20 Treffer. Der vorliegende Bericht wurde auf einer Datenbasis aus dem Zeitraum vom 21. August 2017 bis zum 24. September 2017 erstellt, welcher insgesamt mehr als 5.9 Millionen Einträge von 1534 Nutzerinnen und Nutzern (siehe Appendix A) enthält.

Zunächst beginnen wir mit einem kurzen explorativen Überblick über die verwendete Datenbasis (Kapitel 1.1), auf welchen eine Untersuchung der Unterschiede in den Suchergebnissen (Kapitel 2), ähnlich zum ersten Bericht, folgt, um den Raum für eine mögliche Personalisierung abzustecken. Anschließend wird versucht, über den Login-Status des jeweiligen Nutzers Rückschlüsse über die Unterschiede zwischen den Suchenden zu erklären (Kapitel 2.2). Dies wird bei der vorliegenden Betrachtung als ein möglicher Indikator für eine Personalisierung der Suchergebnisse angesehen. In Kapitel 3 untersuchen wir, wie lange sich eine URL in den Suchergebnissen hält.

1.1 Datensatz und Anzahl der URLs

Die vorliegende Datenbasis umfasst insgesamt ca. 5.9 Mio. URLs, die deutschsprachigen Suchergebnislisten zugeordnet werden können (siehe Appendix A). Die Verteilung auf die einzelnen Suchbegriffe geht dabei aus den folgenden Tabellen hervor. Es ist an dieser Stelle zu beachten, dass die leichten Schwankungen durch die in Appendix A erläuterte Filterung nach deutschen Suchergebnislisten zustande gekommen ist.

Tabelle 3: Gesamtzahl der zu analysierenden Links auf als deutschsprachig erkannten Suchergebnislisten für jeden Suchbegriff.

Person	Anzahl URLs	Prozent
Alexander Gauland	378 720	0.108823
Alice Weidel	389 420	0.111897
Angela Merkel	375 660	0.107943
Cem Özdemir	380 300	0.109277
Christian Lindner	385 700	0.110828
Dietmar Bartsch	398 500	0.114506
Katrin Göring-Eckardt	397 939	0.114345
Martin Schulz	376 540	0.108196
Sahra Wagenknecht	397 380	0.114184
Summe	3 480 159	1.

Partei	Anzahl URLs	Prozent
AfD	375 840	0.152792
Bündnis90/Die Grünen	190 702	0.0775267
CDU	375 680	0.152726
CSU	375 680	0.152726
Die Linke	389 700	0.158426
FDP	376 460	0.153044
SPD	375 760	0.152759
Summe	2 459 822	1.

Tabelle 4 zeigt die Anzahl der unterschiedlichen URLs, jeweils für Parteien und Personen, gegliedert nach der Gesamtzahl der URLs sowie der Anzahl der URLs von eingeloggten Nutzern und nicht-eingeloggten Nutzern. Es wird bereits bei dieser Betrachtung deutlich, dass der Suchbegriff „Bündnis90/Die Grünen“ deutlich heraussticht. Dies ist der Tatsache geschuldet, dass in den Ergebnislisten zu Anfragen nach „Bündnis90/Die Grünen“ eine Vielzahl von Ergebnissen enthalten ist, die in keinem offensichtlichen Zusammenhang mit der Bundestagswahl oder der Partei „Bündnis90/Die Grünen“ stehen und oftmals auch in nicht-deutschsprachigen Ergebnissen resultieren. Daher werden durch unseren Sprachfilter, der nach hauptsächlich deutschen Suchergebnisslisten sucht, viele der Suchergebnislisten aus der Datenbasis herausgefiltert. Wir vermuten, dass dies daran liegt, dass das Plugin ohne Anführungszeichen nach dem Suchergebnis gesucht hat. Daher werden auch Artikel angezeigt, die irgendwo im Text das Wort Bündnis oder ‚grünen‘ beinhalten. Warum dies allerdings viele englischsprachige Links erzeugt, ist unklar.

In jedem Fall werden die Ergebnisse zu diesem Suchbegriff in der späteren Analyse zwar der Vollständigkeit halber manchmal mit berechnet, aber für vergleichende/aggregierte Aussagen nicht herangezogen werden.

Tabelle 4: Anzahl verschiedener URLs (Parteien)

Partei	# URLs gesamt	# URLs eingeloggt	# URLs nicht eingeloggt
AfD	1431	1350	1176
CDU	1282	1195	1058
CSU	823	781	682
Die Linke	1040	855	925
FDP	1106	987	930
SPD	1370	1279	1106
Summe	7052	6447	5877

Tabelle 5: Anzahl verschiedener URLs (Personen)

Person	# URLs gesamt	# URLs eingeloggt	# URLs nicht eingeloggt
Alexander Gauland	791	668	673
Alice Weidel	807	642	708
Angela Merkel	1143	1069	928
Cem Özdemir	725	600	618
Christian Lindner	747	602	649
Dietmar Bartsch	489	424	433
Katrin Göring-Eckardt	767	662	642
Martin Schulz	1071	978	881
Sahra Wagenknecht	540	398	464
Summe	7080	6043	5996

Als Überblick über die Datenbasis haben wir in Abbildung 1 und 2 im zeitlichen Verlauf die Anzahl unterschiedlicher URLs pro Tag für die jeweiligen Suchbegriffe abgetragen. Da die Achsen jeweils normiert sind, lassen sich erste Aussagen über die Diversität der Medien zu den Suchbegriffen treffen. Betrachtet man die Anzahl der unterschiedlichen URLs, die in den Ergebnislisten für die einzelnen Personen oder Parteien enthalten sind, so ist auffällig, dass nur bei der FDP ein Anstieg in der Anzahl der unterschiedlichen URLs zum Ende des Beobachtungszeitraums zu verzeichnen ist. Sonst ist als genereller Trend sichtbar, dass hier eine Abnahme in der Anzahl der URLs zu verzeichnen ist (siehe Abbildung 1). Hierbei ist zu bemerken, dass das grundsätzlich zu beobachtende Abfallen der Anzahl der unterschiedlichen URLs an den letzten beiden Tagen des Analysezeitraums wenigstens teilweise von der am Wochenende geringeren Anzahl von aktiven Nutzern verursacht worden sein könnte. Auf der anderen Seite ist dieser Abfall auch schon am Freitag vor dem Wahlwochenende zu sehen, so dass dies nicht der alleinige Grund sein dürfte.

Abbildung 1: Anzahl unterschiedlicher URLs pro Tag für die Suchen nach Parteien

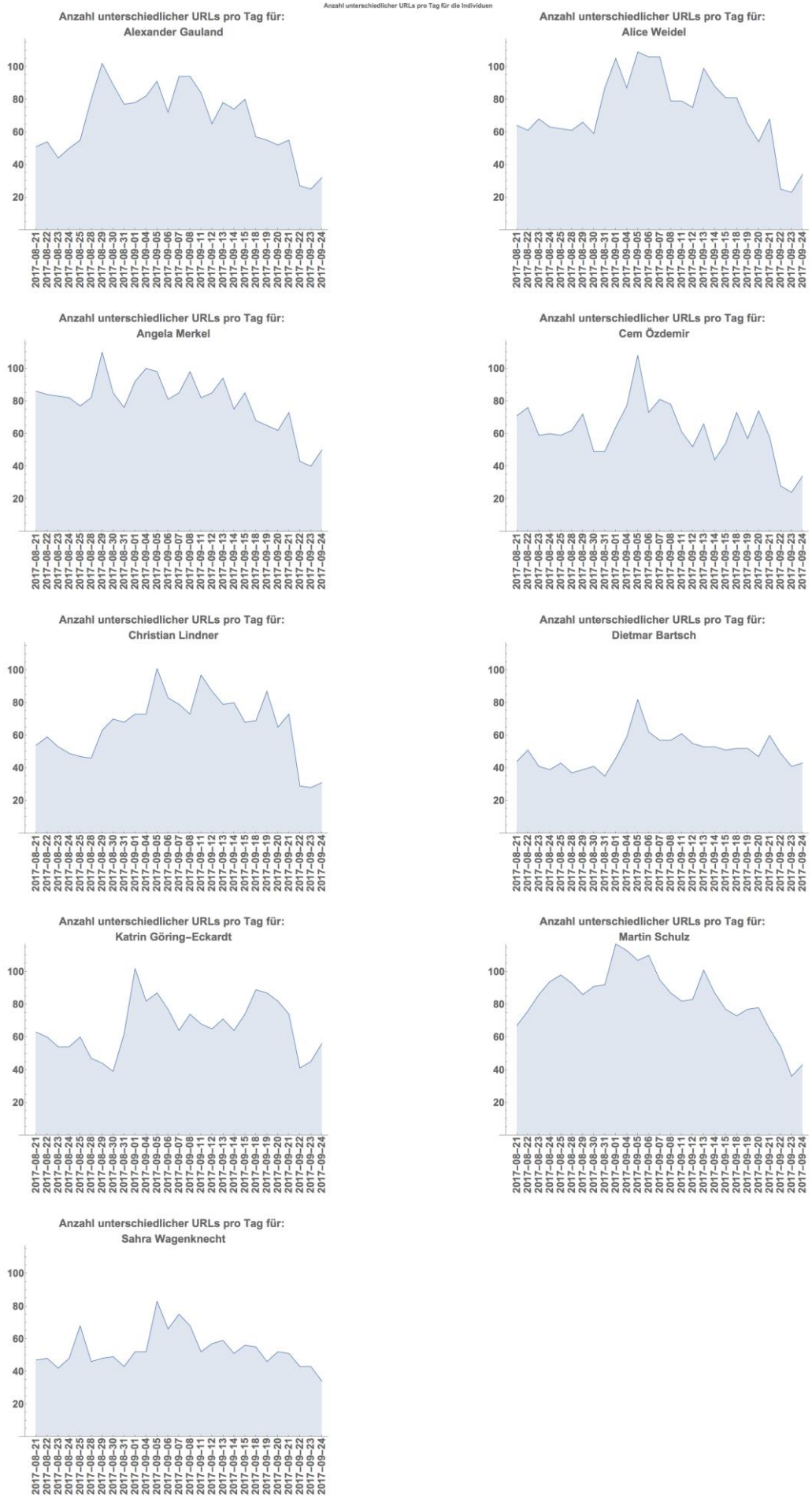


Abbildung 2: Anzahl unterschiedlicher URLs pro Tag für die Suchen nach Personen.

2 Ähnlichkeit der Suchergebnislisten weisen wieder wenig Raum für Personalisierung auf

Die „Ähnlichkeit“ von Suchergebnissen wird in der vorliegenden Untersuchung mit der Anzahl der übereinstimmenden Suchergebnisse für je zwei unterschiedliche Suchergebnislisten gemessen. Nicht berücksichtigt wird bei dieser Untersuchung die Reihenfolge in der die Suchergebnisse dem jeweiligen Nutzer angezeigt werden.

2.1 Gemeinsame Links

Eine erste Frage bei der Analyse der Newsresultate ist, wie unterschiedlich die Ergebnisse sind, die zwei Personen bei der Eingabe des gleichen Suchbegriffs erhalten. Die Erwartung, dass bei Eingabe gleicher Suchbegriffe exakt gleiche Resultate erzielt werden, ist definitiv nicht zutreffend und folgt damit unseren Ergebnissen für die normale Google-Suche im 1. Zwischenbericht (Krafft et al., 2017).

Legt man als Messgröße die Anzahl der gleichen Ergebnisse in der Liste der jeweiligen Suchergebnisse zugrunde, so wäre bei identischen Ergebnisse zu erwarten, dass diese Anzahl stets 20 ist, was der Länge der Ergebnislisten bei der Google Suche auf „news.google.de“, entspricht. Tabelle 6 gibt den Mittelwert für die Anzahl der gleichen Ergebnisse für die jeweiligen Suchergebnisse an, gemessen über alle Paare von Nutzern, die zu demselben Zeitpunkt ihre Ergebnisse erhalten haben. Der größte Wert hier ist mit ca. 16,9 durchschnittlich geteilten Suchergebnissen bei Sarah Wagenknecht festzustellen.

Lässt man die Resultate für „Bündnis90/Die Grünen“ bei der Betrachtung aus den bereits genannten Gründen weg, ergibt sich für die Parteien ein durchschnittlicher Wert von **14,33** durchschnittlich geteilten Links in den Suchergebnislisten, wohin hingegen Suchen nach Personen auf durchschnittlich **15,87** gleiche Beiträge pro Suche kommen. Es zeigt sich also, dass sich die im ersten Zwischenbericht (Krafft et al., 2017) gewonnene Erkenntnis, dass der Spielraum für eine mögliche Personalisierung bei den Parteien größer zu sein scheint als bei den Personen, bestätigt. Insgesamt ist er relativ gesehen aber ähnlich hoch wie bei Googles allgemeiner Suchmaschine.

Tabelle 6: Mittelwert der Anzahl der gemeinsamen Elemente bei der Suche auf Google News nach den entsprechenden Suchbegriffen. Es wurden jeweils die ersten 20 Ergebniseinträge herangezogen. Das bedeutet, dass im Durchschnitt über die Werktage vom 21. August 2017 bis zum 22. September und über das Wochenende vom 23./24. September 2017 zwei Personen, die an demselben Tag nach „Angela Merkel“ gesucht haben, im Durchschnitt 14,3 gleiche Links angezeigt bekamen.

Partei	Gesamt	Eingeloggt	Nicht eingeloggt
CSU	15.7527	15.7585	15.7337
FDP	14.8026	14.8513	14.6675
Die Linke	14.6843	15.2468	13.587
CDU	14.0393	14.0604	13.975
SPD	13.6197	13.6421	13.5511
AfD	13.0992	13.1218	13.0275

Person	Gesamt	Eingeloggt	Nicht eingeloggt
Sahra Wagenknecht	16.9909	17.5412	15.9489
Dietmar Bartsch	16.7312	17.2794	15.7008
Cem Özdemir	16.5472	16.7184	16.1994
Alexander Gauland	16.2691	16.4061	15.953
Christian Lindner	16.1662	16.5241	15.4299
Katrin Göring-Eckardt	15.7706	16.1012	15.1724
Alice Weidel	15.4259	15.8856	14.5333
Martin Schulz	14.5826	14.6178	14.49
Angela Merkel	14.361	14.358	14.3646

Wenn also Nutzerinnen und Nutzer 14-15 gleiche Ergebnisse angezeigt bekommen, beläuft sich der Raum für eine mögliche Personalisierung auf 4-5 Ergebnisse auf der ersten Seite.

2.2 Eingeloggte Nutzer

Nach der Filterblasen-Theorie von Eli Pariser könnte man vermuten, dass Personen, die einen Google-Account haben und eingeloggt sind, grundsätzlich eher mehr Personalisierung erfahren in ihren Suchergebnissen als Personen ohne einen solchen Account bzw. solche, die am Tag der Suche nicht eingeloggt waren. Daher haben wir unsere Analysen noch einmal aufgetrennt in die beiden Gruppen (eingeloggt vs. nicht eingeloggt). Über alle Tage des Analysezeitraumes hinweg waren von den 1500 Nutzern etwas mehr als 2 Drittel eingeloggt.

Bei der genaueren Betrachtung der Werte bezüglich des Status „Eingeloggt“ der Nutzerinnen und Nutzer in Tabelle 6 zeigt sich, dass die Abweichung für Parteien bis auf „Die Linke“ bei unter 0,2 liegt. Bei dieser Partei jedoch erhalten eingeloggte Nutzerinnen und Nutzer 1,5 mehr gleiche Inhalte angezeigt als nicht eingeloggte[1].

Bei den Personen dagegen zieht sich dieser Effekt fast durch alle Suchbegriffe: Durchschnittlich ist die Anzahl der gleichen Suchergebnisse für eingeloggte Nutzer um 0.85 geteilte Links größer als für die nicht eingeloggten Nutzer (Tabelle 7). Für Angela Merkel gibt es allerdings kaum einen Unterschied, auch für Martin Schulz ist er gering. Bei einer ähnlichen Untersuchung, die wir auf den Google-Daten in der allgemeinen Suchmaschine durchgeführt haben (1. Zwischenbericht, Abbildungen 3 und 4) war die Differenz zwischen den eingeloggten und nicht eingeloggten Nutzern auch nicht groß (über die einzelnen Tage des damaligen Analysezeitraumes nur in seltenen Fällen mehr als ein Link), aber fast ausnahmslos zeigte er in die andere, erwartbarere Richtung: Die eingeloggten Nutzer zeigten dort im Durchschnitt meistens eine kleinere Basis gemeinsamer Links als die nicht eingeloggten Nutzer. Auf der News-Seite nun ist dieser Effekt wie beschrieben umgekehrt und bedarf einer weiteren Betrachtung.

[1] Eingeloggt: 15,3; Nicht eingeloggt: 13,7

Tabelle 7: Vergleich der durchschnittlichen Anzahl gemeinsamer Links eingeloggter und nicht eingeloggter Nutzer bei Suchen nach Personen

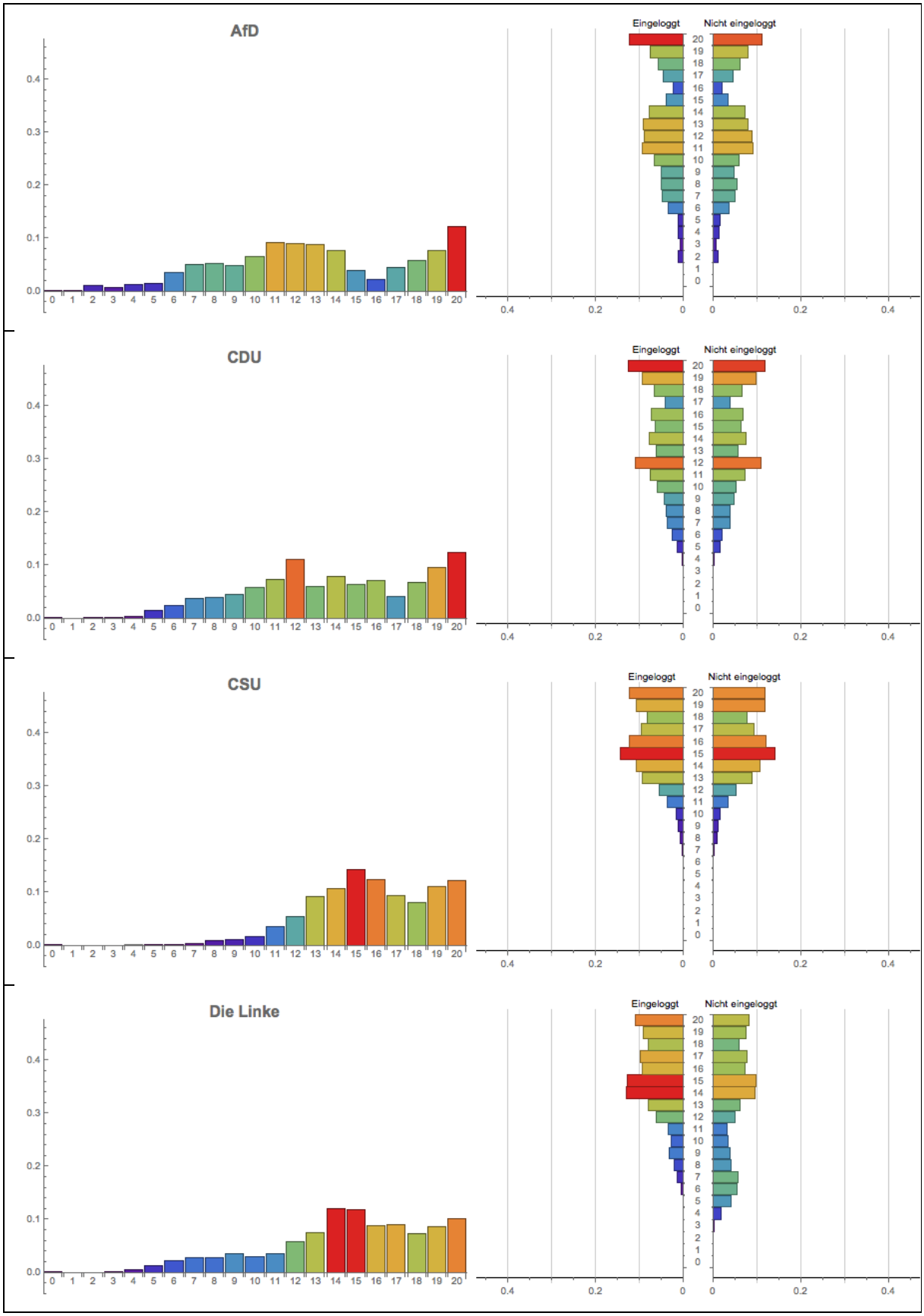
Suchbegriff	Eingeloggt	Nicht eingeloggt	Differenz
Sahra Wagenknecht	17,5412	15,9489	1,5923
Dietmar Bartsch	17,2794	15,7008	1,5786
Alice Weidel	15,8856	14,5333	1,3523
Christian Linder	16,5241	15,4299	1,0942
Katrin Göring-Eckhardt	16,1012	15,1724	0,9288
Cem Özdemir	16,7184	16,1994	0,519
Alexander Gauland	16,4061	15,953	0,4531
Martin Schulz	14,6178	14,49	0,1278
Angela Merkel	14,358	14,3646	-0,0066

Abbildung 3 und Abbildung 4 zeigen die Verteilung der Anzahl an Überschneidungen der Ergebnislisten für die einzelnen Suchbegriffe. Man sieht bei der Betrachtung der Parteien (Abbildung 3) (ohne die Resultate für „Bündnis90/Die Grünen“, s.o.) zwei auffällige Spitzen, jeweils zwischen 11-15 und bei 19-20 geteilten Ergebnissen. Unsere Nutzerinnen und Nutzer scheinen also bevorzugt entweder 11-15 gleiche Ergebnisse angezeigt zu bekommen oder fast vollständig die gleichen Ergebnisse zu sehen. Da dieser Effekt gerade bei der regional ausschließlich in Bayern verwurzelten CSU verschwimmt, könnte es sich um einen Regionalisierungseffekt handeln. Als *Regionalisierung* bezeichnen wir Suchergebnisse, die klar einen regionalen Bezug haben. Bei Parteien sind dies vor allen Dingen Hinweise auf Ortsvereine oder regionale Events. Ein Test dieser Hypothese erfolgt in den folgenden Wochen.

Ein ähnlicher Effekt zeichnet sich auch in der Abbildung 4 bei der Betrachtung der Personen ab. Während fast alle Suchen nach Personen einen deutlichen Spitzenwert bei 15-18 gleichen Ergebnissen zu verzeichnen haben, so liegen bei den beiden Spitzenkandidaten Angela Merkel und Martin Schulz deutlich nach links verschobene Ergebnisse vor mit weniger gleichen Links in den Suchergebnislisten. Hier könnten die großen Wahlkampftouren der Kandidaten in den letzten Wochen vor der Wahl Auslöser für eine höhere regionale Prägung der Ergebnisse sein – auch diese Hypothese werden wir in den folgenden Wochen testen.

Die Aufteilung der Verteilungen in Paare von eingeloggten und Paare von nicht eingeloggten Nutzern zeigt, dass die Verteilung der Anzahl gemeinsamer Links von nicht eingeloggten Nutzerpaaren oft etwas weniger ausgeprägte Maxima hat und einen etwas höheren Anteil an Personenpaaren, die weniger als 10 gemeinsame Links in ihren Suchergebnissen haben. Insgesamt ergeben sich aber keine großen Differenzen, wie auch schon Tabelle 5 nahegelegt hat.

Zusammenfassend lässt sich sagen, dass insgesamt viele Paare von Nutzerinnen und Nutzern ähnliche Suchergebnislisten bekommen. Im **Durchschnitt** über alle Paare sind es für alle Suchbegriffe (außer „Bündnis 90/Die Grünen“, s.o.) mindestens 65% der Suchergebnisse auf der 1. Suchergebnisseite. Die detaillierten Verteilungen in Abbildungen 3 und 4 zeigen aber auch, dass es fast immer auch Paare von Nutzern gibt, die weniger als fünf Links (weniger als 25% der Suchergebnisse auf der ersten Seite) teilen. Diese Nutzer könnten dann durchaus in unterschiedlichen Filterblasen leben – eine hierfür notwendige qualitative Analyse dieser Gruppen steht noch aus.



Mehr zum Projekt finden Sie unter datenspende.algorithmwatch.org

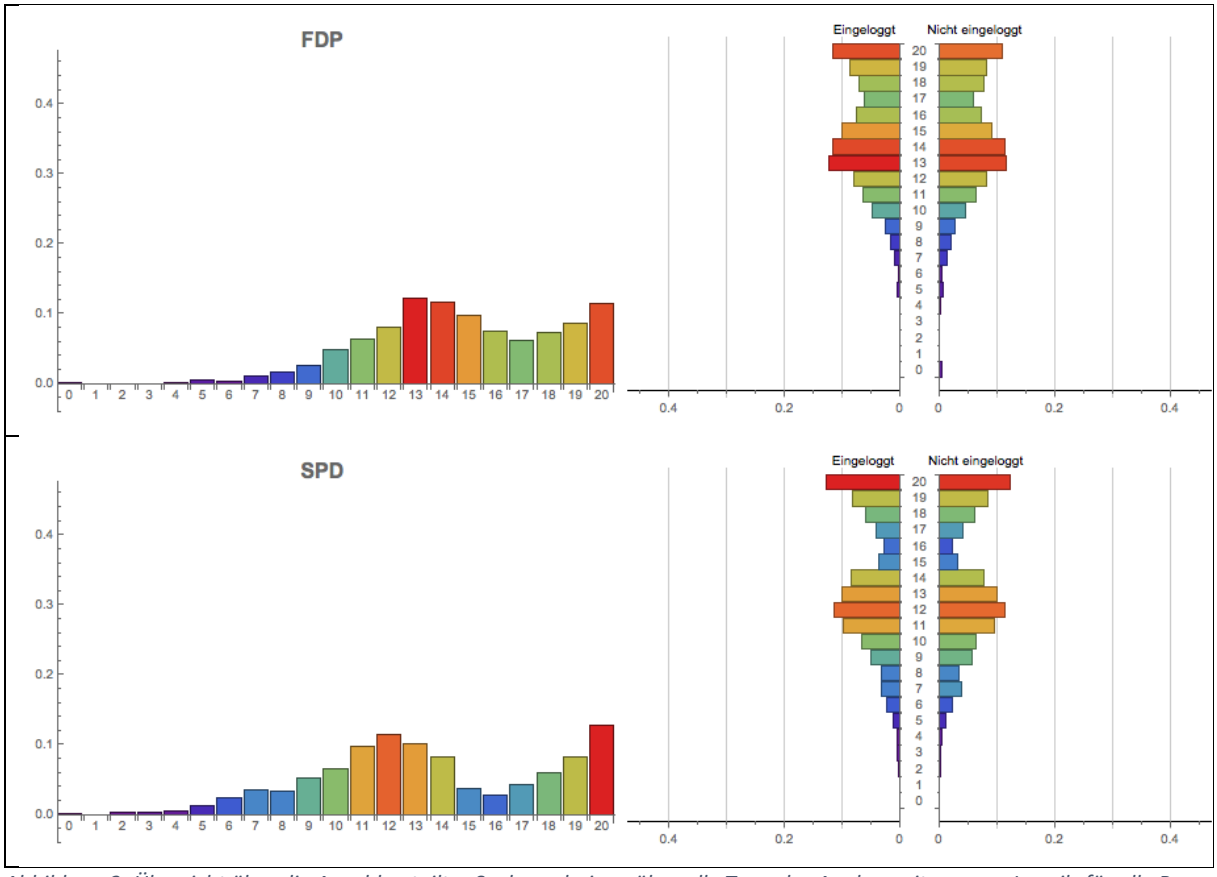
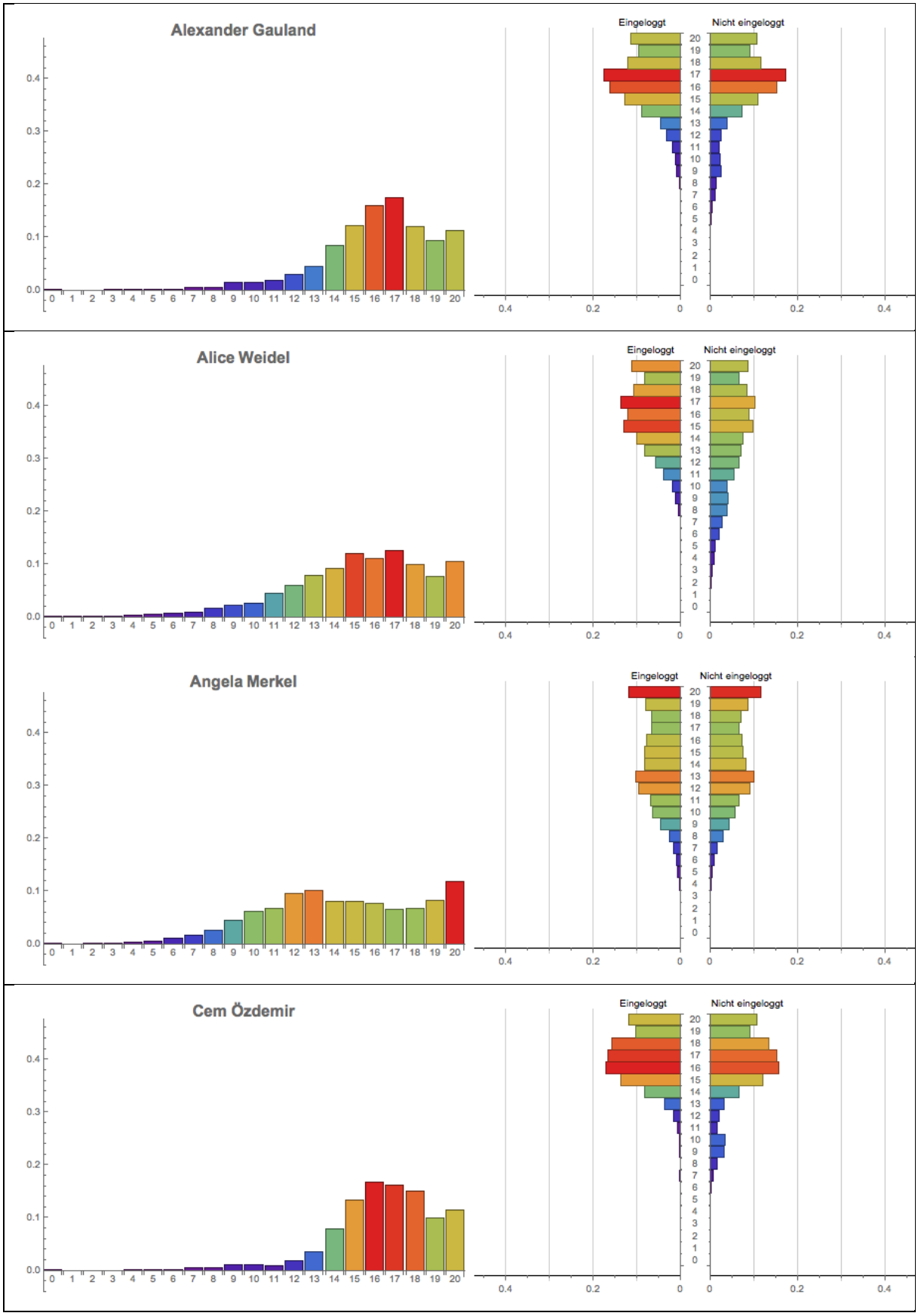
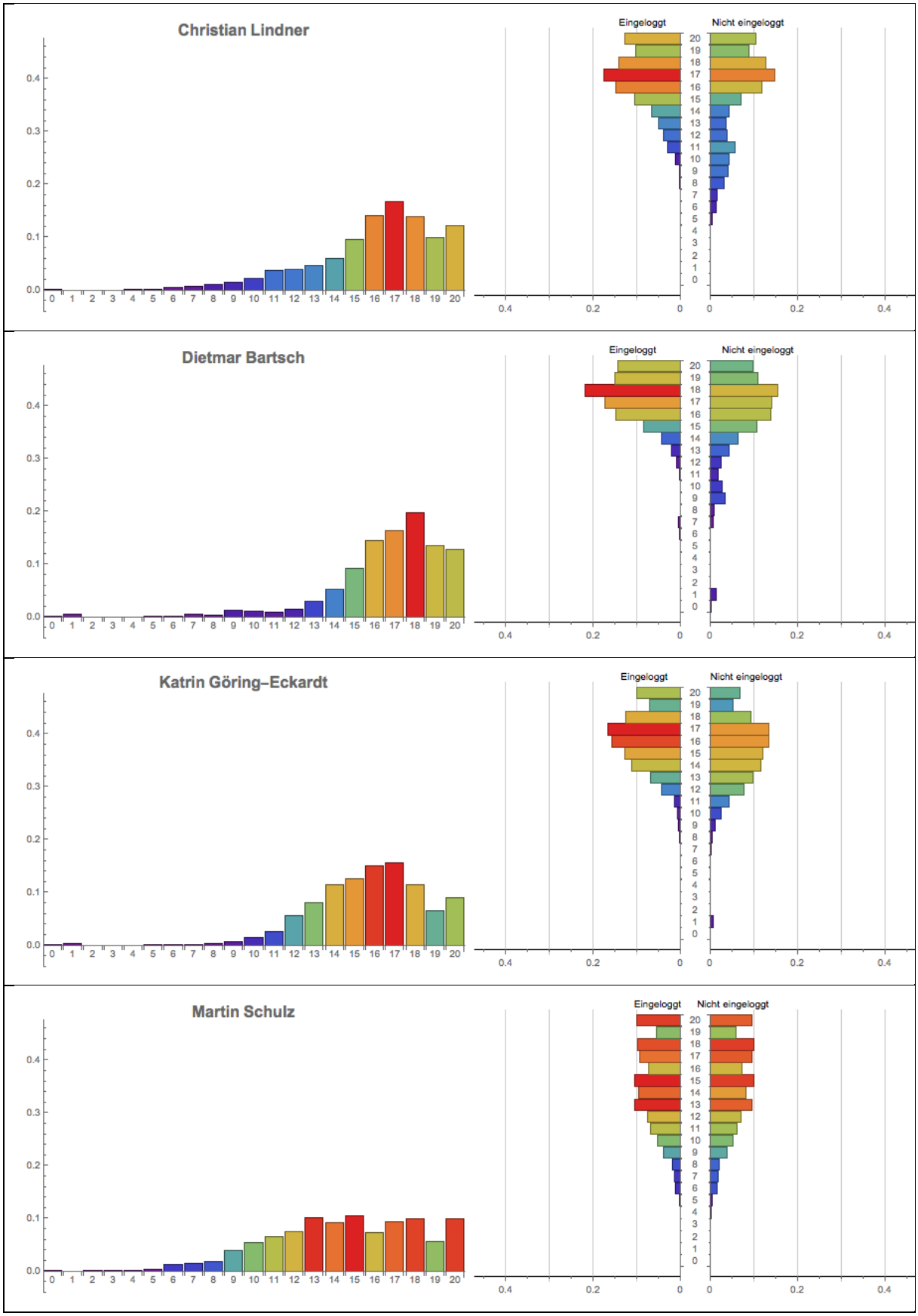


Abbildung 3: Übersicht über die Anzahl geteilter Suchergebnisse, über alle Tage des Analysezeitraumes. Jeweils für alle Paare von Suchanfragen (links) nach einer Partei und noch einmal aufgeteilt in Nutzer, welche in ihren Google Account eingeloggt waren und welche, die keinen haben oder die nicht eingeloggt waren (rechts). Die Farbskala geht von geringer Häufigkeit in blau über grün bis hin zu den Extremwerten in rot, d.h. blaue Säulen zeigen an, dass nur wenige Paare diese Anzahl von gemeinsamen Links haben und rote Säulen zeigen die jeweils höchste Frequenz an gemeinsamen Links unter den Nutzerpaaren an.



Mehr zum Projekt finden Sie unter datenspende.algorithmwatch.org



Mehr zum Projekt finden Sie unter datenspende.algorithmwatch.org

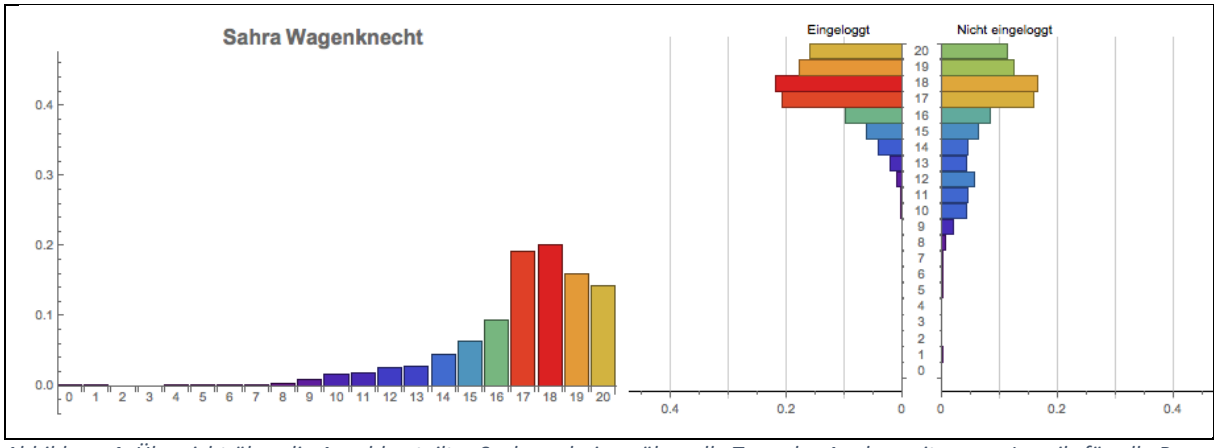


Abbildung 4: Übersicht über die Anzahl geteilter Suchergebnisse, über alle Tage des Analysezeitraums. Jeweils für alle Paare von Suchanfragen (links) nach einer Person und noch einmal aufgeteilt in Nutzer, welche in ihren Google Account eingeloggt waren und welche, die es nicht waren (rechts). Die Farbskala geht von geringer Häufigkeit in blau über grün bis hin zu den Extremwerten in rot, d.h. blaue Säulen zeigen an, dass nur wenige Paare diese Anzahl von gemeinsamen Links haben und rote Säulen zeigen die jeweils höchste Frequenz an gemeinsamen Links unter den Nutzerpaaren an.

3 Lebensdauer der Suchergebnislisten

Schon eine flüchtige Betrachtung der Ergebnisse auf der allgemeinen Suchmaschine und der News-Seite von Google zeigt, dass die Ergebnisse der letzteren deutlich volatiler sind. Während in der allgemeinen Suchmaschine im Wesentlichen auf die jeweiligen Personen- und Parteiwebseiten, allgemeinen Lexika- und Nachrichtensammlungen zur Person oder Partei verwiesen wurden, sind auf den News-Seiten meistens Nachrichten des heutigen und des letzten Tages zu finden. Um diesen Aspekt zu quantifizieren, haben wir neben der quantitativen Auswertung gemeinsamer Resultate in den Ergebnislisten auch die *Lebensdauer* der Resultate betrachtet. Als Lebensdauer wird dabei die längste, zusammenhängende Zeitspanne (definiert über die Zeitpunkte) bezeichnet, in der eine URL für mindestens einen Nutzer sichtbar war.

Genauer definieren wir die Lebensdauer einer URL als die längste Anzahl von konsekutiven Suchzeitpunkten, an denen die URL für (mindestens) eine Person sichtbar ist. Die pro Tag betrachteten Zeitpunkte sind 12 Uhr, 16 Uhr und 20 Uhr für jeden Tag im betrachteten Zeitraum. Die maximale Lebensdauer einer URL beträgt daher 81 Zeitpunkte, da sich die Untersuchung auf insgesamt 27 Tage erstreckt.

3.1 Übersicht über Parteien und Personen

Die mittlere Lebensdauer der URLs findet sich in den folgenden beiden Tabellen, die diese Lebensdauer getrennt für Suchen nach Parteien und Personen darstellen. Die mittlere Lebensdauer der Suchergebnisse bei der Suche nach Parteien ist in Tabelle 8 gezeigt. Auffällig ist hier die insgesamt geringe mittlere Lebensdauer der URLs von ca. einem Tag (pro Tag wurden 3 Suchzeitpunkte herangezogen).

Tabelle 8: Mittlere Lebensdauer der URLs bei Suche nach Partien, angegeben als mittlere Anzahl der Suchzeitpunkte. 3 Suchzeitpunkte entsprechen einem Tag.

Partei	**URL Lebensdauer** **[Anzahl Suchzeitpunkte]**
CSU	4,2
Die Linke	3,9
FDP	3,3
AfD	2,9
CDU	2,9
SPD	2,8

Auf der anderen Seite gab es auch jeweils Links, die sich sehr lange in den Suchergebnissen halten konnten. Hatte es ein Link erst einmal in die Top 10 der am meisten angezeigten Links pro Zeitpunkt geschafft, so blieb er auch deutlich länger in den Suchergebnislisten erhalten. Für diese Analyse wurde pro Zeitpunkt jeweils berechnet, wievielen Nutzern jede URL angezeigt wurde, die mindestens einem Nutzer zu diesem Zeitpunkt angezeigt wurden. Für jeden Zeitpunkt wurden daraus die Top-10 URLs berechnet. Betrachtet man den Mittelwert der Lebensdauer dieser Top-10 URLs, so ist deren mittlere Lebensdauer deutlich höher (Tabelle 9).

Tabelle 9: Mittlere URL Lebensdauer der Top-10 URLs für Parteien, angegeben als die Anzahl der Suchzeitpunkte.

Partei	**URL Lebensdauer [Anzahl Suchzeitpunkte]**
Die Linke	13,4
SPD	10,0
CSU	7,8
FDP	7,0
AfD	5,9
CDU	3,7

Ein ähnliches Bild ergibt sich für die Lebensdauer der URLs bei der Suche nach Personen. Die mittlere Lebensdauer aller angezeigten URLs ist hier deutlich höher als für Parteien, wie in Tabelle111 zu sehen: im Durchschnitt hält sich eine Nachricht ca. 2 Tage. Die mittlere Lebensdauer der Top-10-URLs liegt sogar bei 6-7 Tagen, bis auf bei den beiden Kanzlerkandidaten (3-4 Tage). Die wahrscheinlichste Erklärung liegt darin, dass im Wahlkampf viele Nachrichten über die Parteien nicht immer die Spitzen der Partei erwähnten, während andersherum Nachrichten über die Spitzenpolitiker jeweils auch eine Nennung der jeweiligen Partei enthalten haben dürften. Insgesamt ist also zu vermuten, dass der Pool der Nachrichten über die Parteien zu jedem einzelnen Suchzeitpunkt größer ist als der über deren Spitzenpolitiker, was eine kürzere Verweildauer der Nachrichten über die Parteien nahelegt.

Tabelle10: Mittlere URL-Lebensdauer bei der Suche nach Personen (angegeben in Anzahl der Suchzeitpunkte)

Person	**URL Lebensdauer [Anzahl Suchzeitpunkte]**
Dietmar Bartsch	6,6
Sarah Wagenknecht	6,3
Alice Weidel	5,3
Christian Lindner	5,2
Katrin Göring-Eckhardt	5,0
Cem Özdemir	4,9
Alexander Gauland	4,7
Martin Schulz	4,2
Angela Merkel	3,7

Tabelle111: Mittlere Lebensdauer der Top-URLs bei der Suche nach Personen (angegeben in der Anzahl der Suchzeitpunkte).

Person	URL Lebensdauer [Anzahl Suchzeitpunkte]
Sarah Wagenknecht	24,3
Alice Weidel	22,5
Dietmar Bartsch	20,2
Cem Özdemir	20,2
Christian Lindner	19,1
Alexander Gauland	19,0
Katrin Göring-Eckhardt	18,7
Martin Schulz	11,1
Angela Merkel	9,9

4 Zusammenfassung

Im ersten Zwischenbericht haben wir untersucht, wieviel Raum für Personalisierung bei den Suchergebnislisten auf der allgemeinen Suchmaschine von Google war, in diesem Zwischenbericht haben wir uns dieselbe Frage für die News-Suchmaschine von Google gestellt.

Im Vergleich sind die Suchergebnislisten auf der Google News-Suchmaschine wie erwartet deutlich volatiler, eine URL bleibt im Durchschnitt einen Tag (bei Parteien) und zwei Tage (bei Politiker*innen) in den Suchergebnislisten – auf der allgemeinen Suchmaschine waren dagegen z.B. die entsprechenden Wikipedia-Einträge und die persönlichen Webseiten der Person oder Partei fast immer Teil der Ergebnisliste.
Um die Frage nach der Personalisierung anzugehen, betrachten wir, wieviele der Links auf den Suchergebnislisten zweier beliebiger Nutzer oder Nutzerinnen im Durchschnitt geteilt werden, also beiden Personen angezeigt werden. Ist dieser Wert hoch, bleibt nicht viel Raum für Personalisierung. Bei den Ergebnislisten der allgemeinen Suchmaschine fanden wir im ersten Teil der Analyse heraus, dass bei Suchen nach Politikern davon im Durchschnitt nur 1-2 Links **nicht** von zwei Nutzern geteilt wurden, und bei Parteien 3-4. Von den letzteren waren zudem noch einige Ergebnisse eher regionaler als „personalisierter“ Natur, so dass auch hier nur noch 1-3 Links für die Personalisierung übrig blieben. Wir deuteten das Ergebnis auf der allgemeinen Suchmaschine von Google daher als nicht besonders stark personalisiert, da nur wenig Raum dafür übrig ist. Im Gegensatz zur allgemeinen Suchmaschine werden bei der News-Suchmaschine von Google meistens 20 Suchergebnisse angezeigt statt 9 bis 10, also mindestens doppelt so viele. Unter diesen waren im Durchschnitt ca. 4 nicht geteilte Links in den Suchergebnislisten bei der Suche nach Politikern und bei den Parteien 5-6 nicht geteilten Links – das entspricht also ebenfalls ungefähr einer Verdoppelung. Im Verhältnis bleibt die Anzahl der durchschnittlich nicht geteilten Links also ungefähr gleich.
Trotzdem bleibt festzuhalten, dass es für die meisten Suchbegriffe auch ca. 20-30% der Nutzerpaare gibt, die höchstens die Hälfte ihrer angezeigten Links teilen. Hier bedarf es der qualitativen Untersuchung, worin die Unterschiede genau bestehen und ob sich hier Teilgruppen identifizieren lassen, die

tatsächlich eine sehr unterschiedliche Sichtweise auf die Welt bekommen. Hier wird es dann auch wichtig zu analysieren, ob immer dieselben Nutzerpaare unterschiedliche Suchergebnislisten bekommen, ob diese Personen also dauerhaft in unterschiedlichen Blasen mit wenig Überlappung bleiben. Dies werden wir im letzten Teil des Projektes in Angriff nehmen.

Insgesamt zeigen die bisherigen Ergebnisse, dass der Raum für Personalisierung – ausgehend von Eli Parisers Theorie der algorithmisch erzeugten oder verstärkten Filterblase – eher geringer ist als erwartet. Die Teile der Suchergebnisliste, die nicht mit vielen anderen Nutzern geteilt werden, gilt es jetzt genauer zu untersuchen. Insgesamt zeigt das Forschungsprojekt #Datenspende: Google und die Bundestagswahl 2017, dass es möglich ist, eine Suchmaschine als Black-Box daraufhin zu untersuchen, ob sie stark personalisierte Suchmaschinenalgorithmen verwendet. Eine einmalige Untersuchung kann aber nur zeigen, wie stark dieser Grad zum Zeitpunkt der Erhebung ist. Es scheint daher notwendig, sinnvoll und machbar, eine solche Überwachung auch dauerhaft und automatisiert zu installieren, da eine starke Personalisierung von Suchergebnissen wenigstens potenziell zur Erhärtung von Filterblasen führen kann.

Wir bedanken uns an dieser Stelle bei den Landesmedienanstalten Bayern (BLM), Berlin-Brandenburg (mabb), Hessen (LPR Hessen), Rheinland-Pfalz (LMK), Saarland (LMS) und Sachsen (SLM), die das Projekt fördern, sowie bei Spiegel Online als unserem Medienpartner.

5 Quellen

(Flaxman et al., 2016) Flaxman, S., Goel, S., and Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. Public Opinion Quarterly, 80(S1):298–320.

(Pariser, 2011) Eli Pariser. The Filter Bubble: What the Internet Is Hiding from You. Penguin Press, New York, 2011, ISBN 978-1-59420-300-8.

(Krafft et al., 2017) Tobias D. Krafft, Michael Gamer, Marcel Laessing, Katharina A. Zweig. Filterblase geplatzt? Kaum Raum für Personalisierung bei Google-Suchen zur Bundestagswahl 2017. Technical Report. https://doi.org/10.13140/RG.2.2.29139.07203

Appendix

A Datensatz und Datenvorbereitung

Bei der Erstellung des für diesen Bericht verwendeten Datensatzes wurden die vom Datenspende-Plugin gesammelten Ergebnisse für die Google-Suchen der 25 Werktage vom 21. August 2017 bis zum 22. September 2017 sowie das Wahlwochenende selbst aufbereitet. Im Allgemeinen waren an den Wochenenden weniger Datenspender aktiv, weswegen wir uns auf die Werktage konzentrieren. Das Wochenende direkt vor der Wahl wurde trotzdem mit aufgenommen, um auch die jüngsten Ereignisse vor der Wahl analysieren zu können. Wir haben die Daten so wenig wie möglich gefiltert – aufgrund eines Fehlers in der ersten Version des Firefox-Plugins hatten jedoch alle Nutzer dieselbe ID[2] und suchten bei

[2] Nutzer dieser Plugin Version haben alle unter der ID „{d4e9606e-321a-4fc7-8124-dd0450fdebeb}“ eingereicht.

mit der Spracheinstellung lang=en auf Englisch. Diese wurden daher von der weiteren Analyse ausgeschlossen[3].

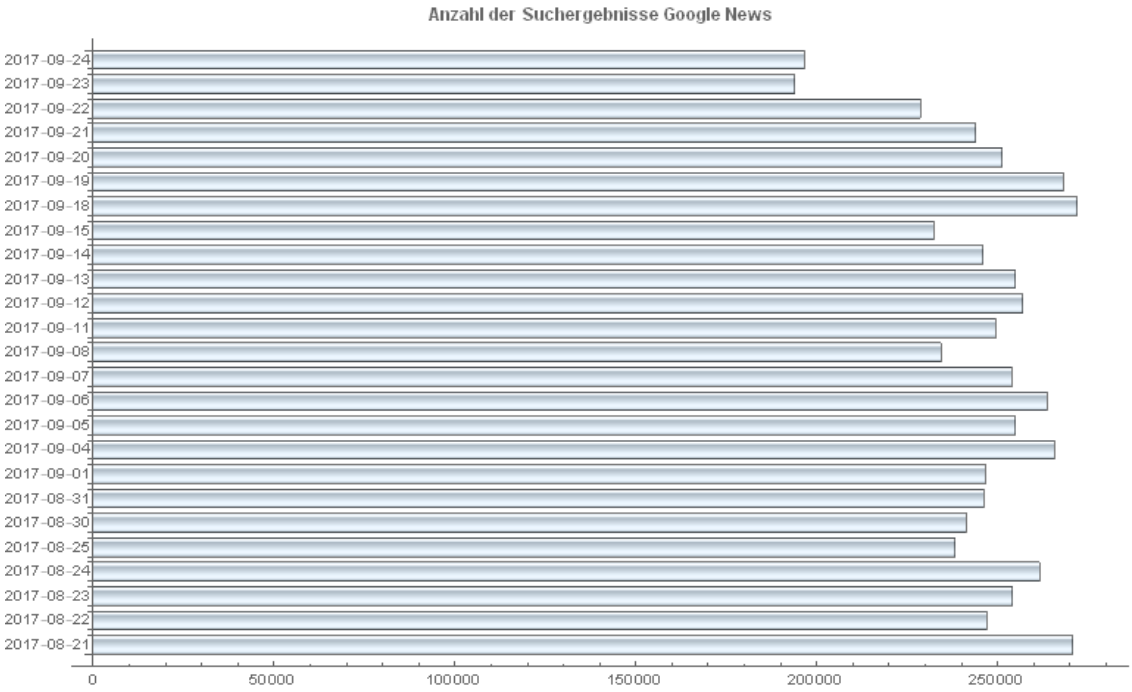


Abbildung 5: Visualisierung der Verteilung der Suchergebnisse auf die Tage des Analysezeitraums.

Die eingesendeten Ergebnisse enthalten – auch wenn alle anderen Nutzer unter der Spracheinstellung lang=de suchten - die unterschiedlichsten Sprachen. Dies könnte z.B. auf Einstellungen des Google-Accounts zurückgehen oder von der IP-Adresse abhängen. Da es um die Bundestagswahl in Deutschland und den möglichen Einfluss von Suchmaschinen auf deutsche Bürger geht, haben wir für die Analysen in diesem Bericht wie für den ersten Zwischenbericht (Krafft et al., 2017) diejenigen Suchergebnislisten herangezogen, die mehrheitlich klar erkennbar deutschsprachig sind. Eine Ergebnisliste wird als deutsch angenommen, wenn mindestens 50% der Einträge den folgenden regulären Ausdruck enthalten:

(.de/ | //de. | //de- | faz.net | handelsblatt.com),

also auf eine deutsche Domain verweisen, mit //de. oder //de- starten (bei Facebook, Twitter und Wikipedia der Fall) oder URLs der beiden deutschen Medien enthalten, deren Domain nicht auf de endet. Tabelle 12 zeigt an einem Beispiel, wie gut die Spracherkennung mit diesem einfachen regulären Ausdruck funktioniert. Versuche mit einem spezifischeren (aber auch willkürlicheren) regulären Ausdruck und einer Sprachdetektionssoftware ergaben keine qualitativ anderen Ergebnisse. Nach Ockham's Razor verwenden wir hier den einfachsten regulären Ausdruck, der seinen Zweck erfüllt.

Tabelle 12 Beispiel für die Anwendung der Sprachenerkennung auf eine Ergebnisliste (066eea906c155260a584acfd77cd1752). Für jede URL wird überprüft, ob der reguläre Ausdruck diese erkennt. Da mehr als 50 % der URLs als deutsch erkannt werden ($\frac{15}{20}$) wird die gesamte Ergebnisliste als deutsch akzeptiert und in die weitere Analyse aufgenommen.

URL	Als deutsch erkannt durch den regulären Ausdruck?

[3] Hierbei handelt es sich um 440 381 von mehr als 6 Millionen Einträgen.

http://www.express.co.uk/news/world/849027/German-election-2017-Angela-Merkel-Die-Linke-Sahra-Wagenknecht	Nein
http://www.rp-online.de/politik/sahra-wagenknecht-russland-sanktionen-schaden-der-wirtschaft-aid-1.7072894	Ja
http://www.augsburger-allgemeine.de/politik/Wagenknecht-und-Bartsch-Ein-Paar-das-alle-ueberrascht-hat-id42636481.html	Ja
https://www.welt.de/politik/deutschland/article168491737/Sahra-Wagenknecht-will-Entsendung-tuerkischer-Imame-verhindern.html	Ja
https://de.sputniknews.com/politik/20170911317385255-deutschland-krim-russland-konflikt-wagenknecht/	Ja
http://www.tagesspiegel.de/themen/reportage/spitzenkandidatin-der-linken-die-verwandlung-der-sahra-wagenknecht/20292204.html	Ja
http://www.bild.de/politik/inland/sahra-wagenknecht/arm-ist-wer-seinen-kindern-kein-eis-kaufen-kann-53150826.bild.html	Ja
https://www.reuters.com/article/us-germany-election-fdp/germanys-fdp-party-leader-cant-imagine-three-way-coalition-idUSKCN1BI34I	Nein
http://www.fr.de/politik/bundestagswahl/sahra-wagenknecht-es-gibt-keinen-gruenen-kapitalismus-a-1347156	Ja
https://www.irishtimes.com/news/world/europe/race-for-bronze-heats-up-in-germany-election-campaign-1.3210512	Nein
http://www.maz-online.de/Nachrichten/Politik/Die-unbekannten-Seiten-der-Sahra-Wagenknecht	Ja
http://www.tagesspiegel.de/politik/linke-spitzenkandidatin-sahra-wagenknecht-ich-hielt-andersdenkende-oft-fuer-idioten/20286898.html	Ja
https://sputniknews.com/art_living/201709051057107249-fashion-club-merkel-wagenknecht/	Nein
https://www.welt.de/politik/deutschland/article168436025/Sahra-Wagenknecht-lehnt-pauschale-Ausgrenzung-der-AfD-ab.html	Ja
https://www.tag24.de/nachrichten/berlin-interview-sahra-wagenknecht-linke-fraktionschefin-bundestag-331730	Ja
https://www.bloomberg.com/news/articles/2017-09-05/ecb-cast-as-villain-for-germany-in-tv-debate-of-smaller-parties	Nein
http://delano.lu/d/detail/news/german-elections-tv-debate-between-small-parties/154600	Nein
http://www.focus.de/politik/experten/voraubauszug-couragiert-gegen-den-strom-wagenknecht-krieg-gegen-den-terror-ist-eine-unglaubliche-heuchelei_id_7537318.html	Ja
http://www.express.de/news/politik-und-wirtschaft/bundestagswahl2017/wahl-o-mat-sahra-wagenknecht--linke--macht-den-test---und-soll-jetzt-das-waehlen-28353776	Ja
http://www.ostexperte.de/sahra-wagenknecht-interview	Ja

Insgesamt sind bei den Suchergebnissen in Google News ca. 5,4 Mio. deutsche Resultate und ca. 0,7 Mio. internationale Datensätze enthalten, welche sich wie in Abbildung 6 gezeigt auf die Tage des Analysezeitraumes verteilen.

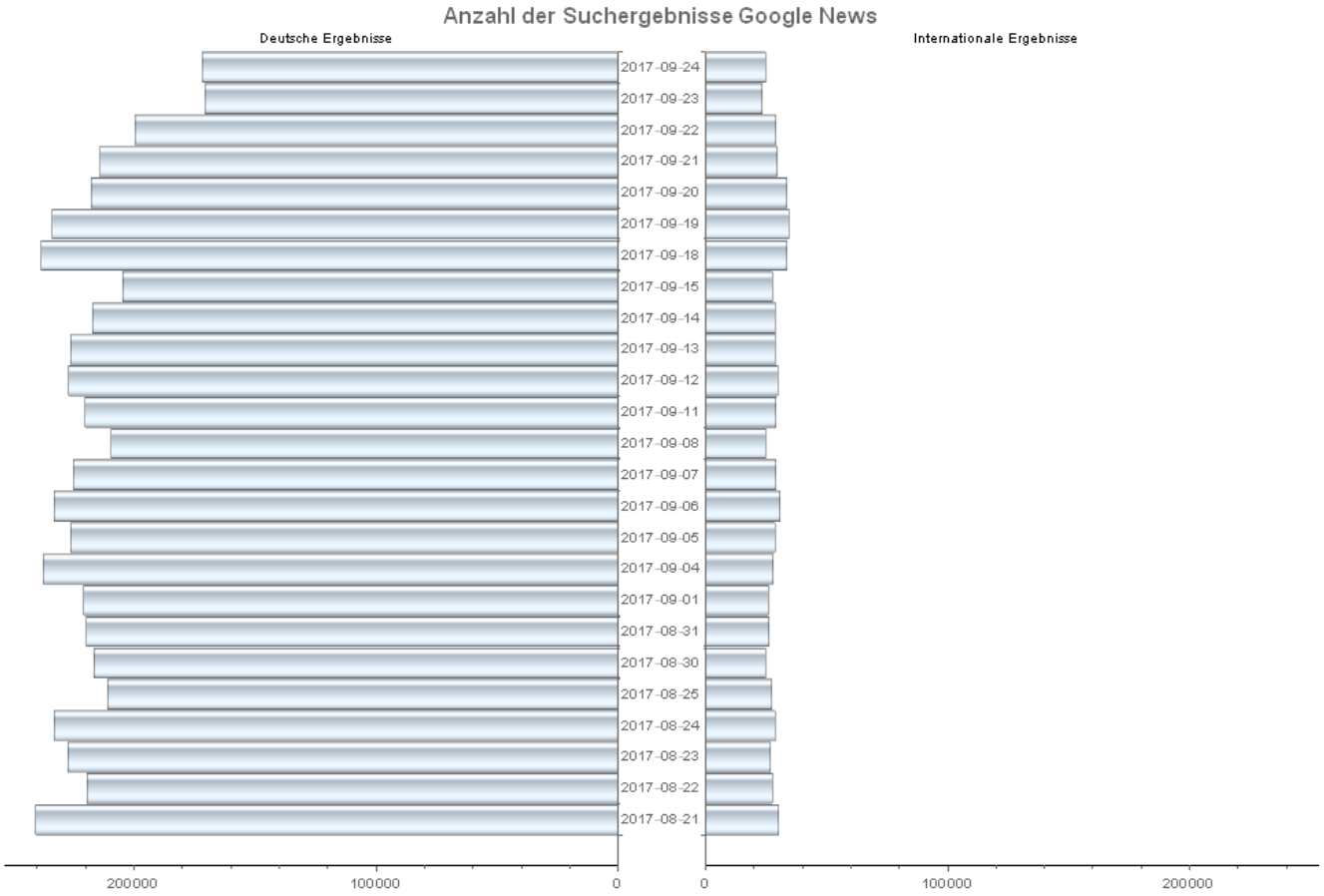


Abbildung 6: Anzahl der Suchergebnisse bei Google News, jeweils aufgetrennt nach deutschen und internationalen.