Can Web Browsing Histories Predict Users' Attitudes Towards Immigrants, Climate Change, and Democracy?

Nora Kirkizh, Roberto Ulloa, Jürgen Pfeffer, Sebastian Stier
Department of Computational Social Science, GESIS, Cologne, Germany
Technical University of Munich, The School of Governance, Munich, Germany
{roberto.ulloa, sebastian.stier}@gesis.org
{eleonora.kirkiza, juergen.freffer}@tum.de

Abstract

Browsing behavior and visits to websites such as donation platforms, social media, streaming service providers, and even online gambling can reflect individuals' life-style, while, as research shows, life-style itself is a predictor of individuals' political issue preferences and attitudes. In this paper, we linked 19,000,000 website visits generated from web tracking of 1,000 users in Germany to self-reported political attitudes to investigate whether website choices can predict individuals' political preferences. Our best performing machine learning algorithm, random forest, was best at predicting individuals' interest in politics, and democratic attitudes. The web browsing histories could not predict individuals perceptions of climate change and populist attitudes, and only partially signal perceptions of immigration. By showing the potential of web browsing data to reveal individuals' political orientations such as authoritarian or democratic attitudes, our cross-validated evidence has normative implications for political campaigns, online privacy, and democracy in general. This study also makes methodological contribution to the literature on linking political preferences to online behavioral data.

Introduction

For political parties, individuals who have particular attitudes towards specific policies but did not vote are valuable to target with political messages. Since people often prefer to read and hear information that confirms their existing believes Kunda (1990) and if political actors know what people believe in by monitoring their web-traffic, they can much better create targeted adds that comply with voters' believe system, and therefore, increase the manipulative power of advertising. Research has already shown that right-leaning websites are tracking their audience more extensively than left-leaning websites Agarwal et al. (2020), while individuals prefer to keep their political views or attitudes private.[1][2]

[1]https://policies.google.com/privacy/key-terms#toc-terms-sensitive-info

[2]https://www.propublica.org/article/how-companies-have-assembled-political-profiles-for-millions-of-internet-us

Research has shown that based on Facebook likes models can predict if a person is Democrat or Republican Kosinski, Stillwell, and Graepel (2013), or even vote choice Cerina and Duch (2020). Visits to untrustworthy news websites are related to people's populist attitudes authors (a) and party affiliation Guess, Nyhan, and Reifler (2020). In this study, we argue that in the time of rising political polarization, voters of specific political views might be identifiable based on their web browsing behavior. Building on previous work that found a strong relation between lifestyles and political orientations DellaPosta, Shi, and Macy (2015), we test this argument based on three-month web browsing histories from 1,000 individuals living in Germany. We examine if individuals' political attitudes are identifiable from *general* website choices, not just their news-related behavior that is the focus of most previous research. Overall, we analyze 19,026,887 website visits and the associated URLs. Prior to the tracking, we asked our panelists to answer survey questions measuring their attitudes towards *(1) immigration, (2) democracy, (3) climate change policies, (4) trust in public institutions, and (5) populist attitudes.* — policy dimensions that reflect manifestos of major political parties in Germany, and parties' Facebook pages authors (b). We also measured panelists' interest in politics, and attitudes towards the European Union (EU). We chose Germany because the country represents a state, whose political landscape has been experiencing a rise of radical right populist parties, which are challenging the norms of democracy. In the challenging for democracy political context, the studies of potential weaknesses in voters' privacy protection on the Internet become even more important. We found that from web browsing histories, the best performing algorithm is random forest and it was able to detect signals about individuals' attitudes towards democracy, and interest in politics. The web browsing histories are not informative to identify attitudes towards immigrants (except perceptions of Islam), populism, and climate change. Trust in public institutions such as parliament are also not distinguishable from zero. Consistent with our theory on the connection between life-style related websites and political orientation, our predictions improve significantly when we apply visit duration thresholds to keep only websites, where people spend considerable amount of time.

The contribution of this paper lies in several dimensions. First, we find that sensitive information about' political attitudes can be revealed from individuals' browsing histories. Hence, further development of digital privacy policies should consider that this knowledge could be potentially used by ads distributors like Google Englehardt and Narayanan (2016) and their customers such as political actors to target voters without asking them about their attitudes towards specific policies directly. Coupled with the fact that these processes are hidden to citizens and the general public, the findings are troubling for democracy. Second, predictability of attitudes towards democracy from visits to life-style related websites suggests that political polarization goes beyond political news consumption. Hence, tracking differences in website preferences between democratic and authoritarian individuals over time can potentially be used as a measure of dynamic polarization at scale. Third, from a methodological perspective, in this paper we introduce an approach on how to draw classification from domains that are relevant for social science research questions, and that measurement of peoples' attitudes based on web tracking data is compatible to self-reported attitudes. Finally, the paper contributes to a growing body of research showing that digital traces can be used for inferences about people's political and social attitudes Cerina and Duch (2020); Kerna et al. (2019); Stachl et al. (2020); Shi et al. (2017) by showing that website choices can be stronger predictors of individuals' political orientations than social media because website visits reflect peoples' life-style.

Related work

A number of studies have used digital footprints, — Facebook likes, smartphone or website logs — to identify peoples' personality traits and other attributes Lambiotte and Kosinski (2014); Settanni, Azucar, and Marengo (2018). Here we provide an overview of the recent literature on using digital trace data to learn about peoples' political tastes, and major studies on association of political attitudes with individuals' lifestyle.

Predicting Individuals' Attributes from Online Behavior Patterns

Social media Social media data can help to find individuals' perfect job match Kerna et al. (2019), to measure emotions Kramer, Guillory, and Hancock (2014) or personality traits. Patterns of Facebook usage such as number of friends, followed groups or published photos can predict big five personality traits like openness, extraversion, and agreeableness among others Evans, Gosling, and Carroll (2008); Golbeck, Robles, and Turner (2011); Bachrach et al. (2012). Facebook likes can predict individuals' attributes and life-style characteristics including alcohol drinking, religion, and even relationship types Kosinski, Stillwell, and Graepel (2013).

Web browsing logs However, social media data might display socially desirable picture of individuals because people do not want the public to know about their true interests. Self-reported website choices, however, might be more revealing and consistent with peoples' personalities. Websites choices can predict personality traits Kosinski et al. (2012). For instance, high level of emotionality correlates with visits to websites related to sports, while calm personality is associated with photography. Observational data with websites extracted from Facebook likes showed that self-reported website choices are robust in predicting peoples' personalities Kosinski et al. (2014). Collected from smartphones data can reveal personalities as well. For instance, phone activity correlates with extraversion, and music apps with openness Stachl et al. (2020). Individuals' browsing behavior is also capable to predict demographics features. Age and gender are the most predictable from browsing behavior Hu et al. (2007); Kosinski, Stillwell, and Graepel (2013) setting a benchmark for other attributes.

Digital trace data and political attributes The review of related literature shows that researchers has been focusing on personality traits and demographic attributes of individuals while political orientations remain understudied. Some studies include politics-related attributes into their prediction models as a secondary variables of interest. For example, using Facebook likes, machine learning models can predict individuals' vote choice Cerina and Duch (2020), whether a user is a Democrat or Republican Kosinski, Stillwell, and Graepel (2013), while visits to untrustworthy websites are associated with populist attitudes authors (a) and party affiliation Guess, Nyhan, and Reifler (2020). In this paper, we offer a direct prediction of broader set of individuals' political attitudes based on their browsing histories.

Political Attitudes, Political Behavior and Lifestyle

Political attitudes are based on a set of beliefs with which individuals approach political issues. Attitudes are much more stable than opinions, and they are hard to shift Krosnick (1991), which makes them a good predictor of individuals' political ideology or even vote choice. In fact, political attitudes often become a source of decision-making and political behavior like voting Dalton (2000). For instance, in Europe, anti-immigration attitudes are a predictor of voting for radical-right parties Rooduijn (2018), attitudes towards climate change policies are associated with voting for green parties Harteveld, Kokkonen, and Dahlberg (2017). Research also demonstrates that people who have never voted before often rely on their political attitudes when voting Arcuri et al. (2008). Hence, political parties or candidates, when drafting their policy platforms, often appeal to voters' political attitudes. This makes identification of political attitudes especially desirable for political parties or candidates while individuals would prefer their attitudes towards, for instance immigration or foreign workers, remain private.

In this paper we attempt to predict political attitudes from individuals' web browsing behavior. However, can website choices signify about political attitudes and what are the mechanisms? We rely on existing literature, which demonstrates that people with distinct lifestyle preferences also have specific political attitudes and personality DellaPosta, Shi, and Macy (2015); Gerber et al. (2010); Fatke (2017);

Shi et al. (2017).[3] For example, immigration attitudes can be linked to lack of compassion or traveling Klimecki, Vétois, and Sander (2020), negative attitudes towards climate change policies and support of authoritarian policies to deal with crime are associated with low level of generalized trust Gauchat (2018); Lo Iacono (2019), populist inclinations can be the result of an individual being in debt or loosing job Wiedemann (2020), or low level of agreeableness Assuming that offline behavior is generally reflected in online behavior, each of the suggested associations might be derived from visits to specific websites. For instance, hotel and flights booking platforms can be a proxy of cosmopolitan vs. nationalist orientation, whereas gambling and charity websites can be a proxy of individuals with high level of empathy. However, people choose websites according to their predispositions. We build on the theory of selective exposure Garrett (2009); Prior (2013), implying that individuals choose political information based on their existing world view. We apply this theory to browsing behavior assuming that individuals choose websites based on their preferences and attitudes.

Data and Measurement

In this paper, we use two types of data: web browsing logs and online survey responses. The data was collected with approval of the Oxford Internet Institute's Departmental Research Ethics Committee at the University of Oxford (Reference Number SSH IREC 18 004).

Web Tracking

We acquire web browsing histories of respondents from an online access panel maintained by Netquest, a market research company. Respondents enter the panel only via invitation, which they receive via e-mail under user consent. The panel consists of respondents who agreed to install plugins tracking their web browsing behavior on desktop computers. If they consent to participate, panelists receive additional incentives in case they do not stop the tracking for longer than seven days. Participants have the possibility to pause the tracking tools at any time. The tracking tools would then be interrupted for 15 minutes. Personally identifiable information is algorithmically anonymized by Netquest. We utilize web browsing histories from 1,003 panelists living Germany. The tracking period is between mid-March and mid-June, 2019. The dataset includes anonymized IDs, visited URLs, domains, and time spent on the web page. Overall, the dataset comprises 19,026,887 URLs, and 96,093 unique domains. Table 1 shows the distribution for mean duration and number of visits per panelist and other main summary statistics.

We also tested if our data represent behavior of the general population. Since our panelists were aware of the tracking, they might have altered their behavior. For instance, they might have started to visit more news websites to learn more about political issues or the other way around; in short, respondents who are being tracked might be more careful in revealing their political interests and inclinations. We correlated visits made by our panel to media websites with ground truth data on the visits of the German general population recorded by the "Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V." (IVW). The correlation between the ranking of news sites visited by our German tracking panelists and the IVW data is strong ($\rho = 0.73$). These evaluations give us confidence that our tracking data provides a fairly accurate representation of visited websites by internet users in Germany.

Table 1: Descriptive statistics of web tracking variables. There were 1,000 panelists, 19,026,887 URLs, and 96,093 unique domains.

Statistic	Mean	St. Dev.	Min	Max
N visited URLs	18,080.07	23,864.05	53	191,526
Duration (sec.)	469,971.90	539,768.10	569	4,660,753
N unique domains	362.04	328.97	9	2,279
μ visits/domain	43.36	37.30	3.28	376.76
μ dur/domain (sec.)	1,373.64	1,955.56	56.90	44,116.23
μ dur/URL (sec.)	33.34	23.58	1.62	276.12

In addition, following Guess, Nyhan, and Reifler (2020) we evaluate to what extent privacy attitudes of tracking panelists diverge from panelists who participate in surveys but do not have tracking tools installed. To identify a potential "opt in bias", we implemented the same privacy attitude battery in a sample of German participants drawn from the regular online access panel of Netquest without web tracking. In total, 1,00o participants were sampled according to German population margins for gender, age and education. Respondents were presented the following statements and asked about their (dis)agreement on a five-point scale: Personalized advertising makes me afraid; I am concerned about how much data there is about me on the Internet; and My privacy on the Internet does not matter to me. A figure with privacy attitudes online are uploaded on the project repository.[4] The figure shows that there are minor differences in the privacy attitudes of online panelists with and without web tracking technology installed, which brings the results of this paper closer to generalizability.

Survey

To combine digital trace data with panelists' responses authors (c), we surveyed the study participants parallel to the web tracking. To infer their political attitudes, panelists were asked a set of questions related to diverse political issues: *immigration*, *democracy*, *climate change*, *trust to public institutions*, *populist attitudes* as well as *political ideology* on the left-right scale, political interest and attitudes towards the EU integration. The responses are placed on Likert (from strongly disagree to strongly agree) or 1-11 scales among others. In addition, we asked demographic questions such as age, gender, education, and income. We used the question wording established by prominent annual survey panels such as Eurobarometer, European Social Survey, and World

[3] https://www.pewresearch.org/politics/2014/06/12/section-3-political-polarization-and-personal-life/

[4] Link to OSF repository.

Values Survey. The table and a figure with summary of survey items is available on the project repository.[5] Differences in the number of questions per dimension represent the heterogeneous underlying concept structures. For instance, the dimension on immigration requires to ask about attitudes towards immigrants and their association with crime as well as jobs, while climate change attitudes can be measured with only one question, aiming to identify acceptance or rejection of climate change.

Methodology

In the analysis we use domains as a unit of observation that will be predicting political attitudes. We choose domains instead of URLs because similar domains appear in the data set much more often across individuals' browsing histories than URLs, i.e., the use of URL would produce a very sparse representation. We use two different thresholds for further analysis of the web tracking data: domain visits with duration one and five minutes. These thresholds are important because based on theory, we are interested in website visits that can signify the lifestyle of an individual. Domains where individuals spent more than five minutes are more likely deliberately visited. We choose the one minute cutoff as an additional robustness test. After we removed "short" domain visits, where individuals spent less than one minute, the data generates 1,632,769 visits (35,380 unique domains) with 1,003 respondents.

Method I: Machine learning models with pre-existing domain categories

As a method of dimensionality reduction we applied existing domain categories from Webshrinker. With categories from Webshrinker, we were able to categorize 13,824 out of 35,380 unique domains in the dataset on which respondents spent at least 1 minute, and the domain itself had at least 5 visits. The domains fall into the following categories: search engines, blogs, dating, gambling, social media, travel, news, games, health among others (full list of categories is available on the paper's OSF repository. Figure 1 shows an example of the association between selected attitude: satisfaction with democracy and social networking (social media). Plots for the rest of attitudes and domain categories are available on OSF repository. The major concern about the distributions of attitudes of interest is that they are ordinal variables which gives less information to the machine learning algorithm to identify an association between two variables. Panel A of Figure 1 illustrates that predicting variable which is the number of visits to a domain category "social networking" is severely skewed towards zero because not many respondents visit social media websites. We apply a log transformation to the x-axis and find that the the slope of the linear function went down. This demonstrate how important the model specification is to estimate the association between variables correctly. We therefore use three different algorithms to test the predictability of website choices (visits to websites), we exploit the base line model, where we estimate the average

[5]Link to OSF repository.

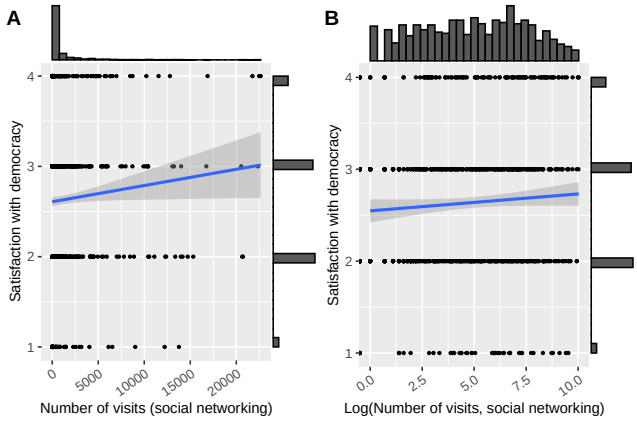

Figure 1: Association of visits to social media websites and satisfaction with democracy with distributions.

predictability power from a training dataset, elastic net regression, which is sensitive to multicollinearity among other advantages, and random forest, which identifies variables with the largest explanatory power. To demonstrate if any of the chosen algorithms are working, we compare our estimates with benchmark socio-demograpgics such as gender, income, education, and age.

The basic regression model looks as follows:

$$Political\_Attitude = \alpha + \beta_1 C_1 + \beta_2 C_2 + ... + \beta_n C_n, \quad (1)$$

where $Political\_Attitude$ represents survey-based attitudes, α is the intercept, and β is a regression coefficient for every domain category C. Overall, we have 16 questions measuring political attitudes, which we include into the model one by one.

For each model, we measured its ability to predict new cases using 10-fold cross-validation, i.e., we ran 10 repetitions of the cross-validation process while randomizing the selection of the 10 folds each time. Each 10-fold cross-validation repetition splits the data in 10 fixed parts and uses 9 to predict the 10th. Repeating the cross-validation ensures that the prediction was not an artifact of the selection of the 10 fixed parts. We only considered a dependent variable to be "predictable" for a given k, if the prediction was statistically significant ($p < 0.05$) in all the cases in which the cross validation was repeated (i.e., 10 out of 10). The statistical significance was calculated using Pearson correlations between the predictions and the dependent variables on the test splits.

Method II: Machine learning model with domains clustered by Singular Value Decomposition

To predict political attitudes, we follow the methodology used by Kosinski et al. Kosinski et al. (2016) to predict personality traits with Singular Value Decomposition (SVD) based on likes to Facebook pages Golub and Reinsch. We assume that analyzing web tracking data poses a similar scenario because (1) both of them can be considered digital traces, and (2) a visit to a website indicates interest in a

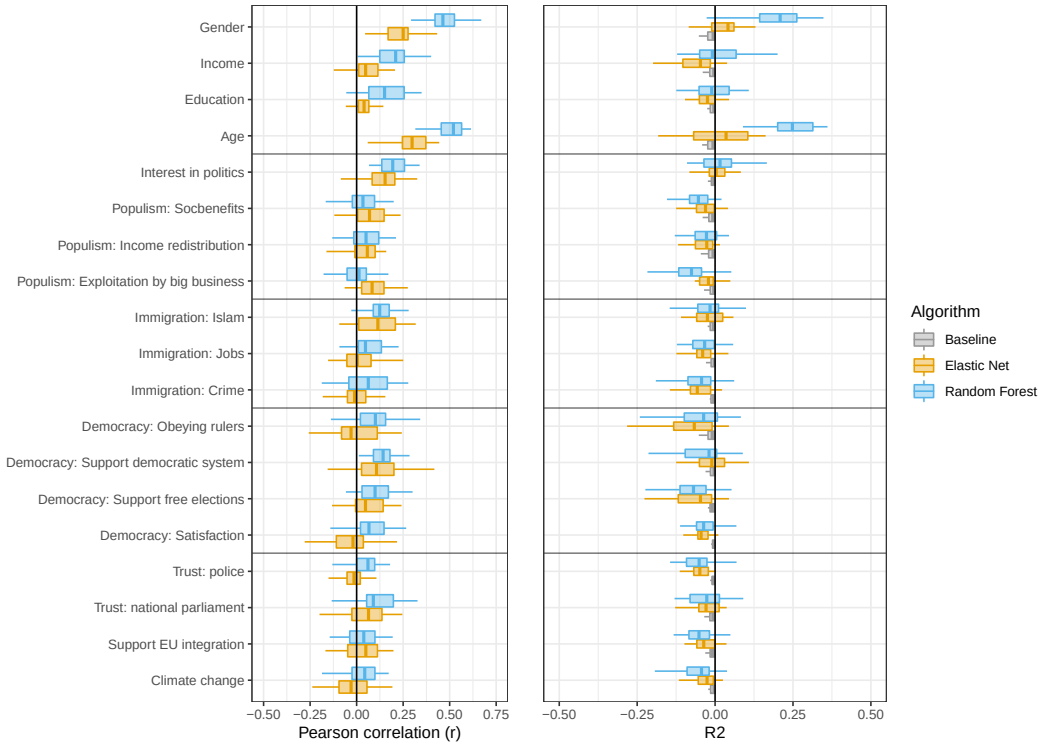

Figure 2: Box and whisker plot of prediction performance measures from repeated cross-validation for each political attitude and socio-demographics. The middle symbol represents the median, boxes include values between the 25 and 75% quantiles, and whiskers extend to the 2.5 and 97.5% quantiles.

similar way that a like to a Facebook page does. As opposed to Facebook likes, however, visits to websites are not binary and also contain a duration. Below, we explain the adjustments taken to deal with these two features.

Pre-processing. We represent the browsing data in a matrix, where rows are individuals (respondents) and columns are unique domains. Each cell of the matrix contains the number of visits that a participant made to a particular domain. Additionally to the initial 5min/1min cutoffs explained before, we trim the matrix by excluding domains that were visited by less than 10 respondents and respondents that visited less than 10 domains. This procedure removes rare domains and occasional participants that do not contribute a lot of information to the model considering that our relatively small sample size would not represent these visits properly.

Decreasing the number of dimensions. Since our respondent-domain matrix is sparse, we apply Singular Value Decomposition (SVD), a popular dimensionality reduction approach based on eigendecomposition Golub and Reinsch. Given the sparsity of our matrix, SVD also helps representing the data in a more compact dimensional space. We preferred SVD over other methods as it is computationally efficient and suitable for our exploration.[6]

[6]LDA, suggested in Kosinski et al. (2016), offered small improvements to predictions at high, however, computational costs.

SVD "dimensions" (i.e., left and singular values) can be interpreted directly if they are properly rotated (using varimax rotation). Each dimension then will be represented by all domains and a "coefficient" for each them that can be used to sort the domains according to their importance inside each dimension, also taking consideration that such coefficients could be negative values. Thus, SVD allows to identify domains with large explanatory power or domains that are the most relevant for the prediction of political attitudes.

One caveat with any dimensionality reduction approach is the selection of k; using a scree plot of the explained variance, a good k is often selected by visually identifying the "elbow", i.e., the point after which adding more dimensions does not decrease the explained variance. However, given the exploratory nature of this study and the difficulty of the task, we identify the best k by training and testing models for each k in a comprehensive set of values ranging from 5 to 500: $\{5\} \cup \{10\text{i}: 1 \leq \text{i} \leq 20\} \cup \{25\text{i}: 8 < \text{i} \leq 12\ \} \cup \{50\text{i}: 6 < \text{i} \leq 10\ \}$; $k \in Z$.

Results

The regression analysis focuses on five dimensions of political attitudes: immigration, democracy, climate change, populism, trust (as well as interest in politics in general). We first report prediction performance of the models and domain categories ranked by importance in the model (how much variance the domain category explains in the model).

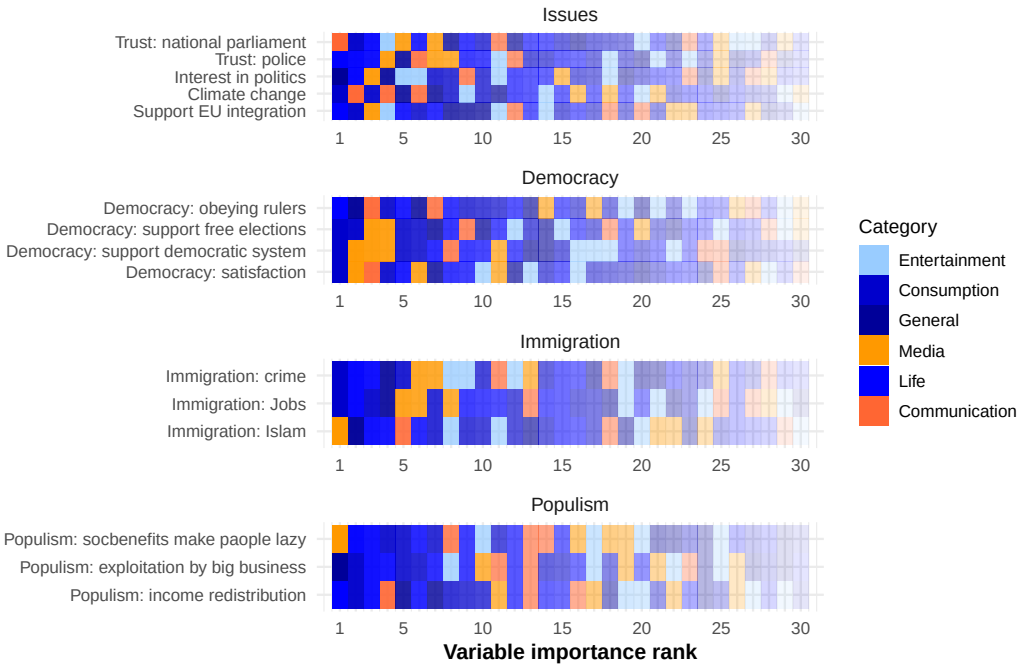

Figure 3: Domain categories ranked by importance in the model for attitudes towards policy issues, democracy, immigration, and populism. The fading effect on the plot represent the decrease in the importance of each domain category since top five domains explain the most of variance. Color represents two palettes — orange and blue — in order to distinguish between domains related to media/communication and consumption/life style. The interactive version of the plot to see the specific domains behind each color-coded cell is published on the project repository.

Second, we apply SVD to the domain level analysis. And third, we use SVD components (dimensions) as predicting variables.

Models' prediction performance. We first predict the political attitudes with domain categories and several machine learning algorithms described in Method II. Figure 2 reports shows the performance of baseline models, elastic net, and random forest. Across all political attitudes, random forest performed best. However, even with the most sophisticated algorithm, Pearson correlation coefficient (r) is significant (within 2.5 and 97.5% quantiles) only for a few political variables: interest in politics, and attitudes towards democracy. Although significant, the correlation coefficients are small: median r=0.15 for interest in politics, r=0.13 for support democratic system. Random forest and elastic net models were also able (within 25 and 75% quantiles) to signal populist attitudes, attitudes towards Islam, support for free elections, satisfaction with democracy, and trust in national parliament. The R2, however, is either very small or even negative. Random forest and elastic net, however, are better at predicting socio-demographic covariates than political attitudes.

In Figure 3 shows variable importance rank for each domain category in the model, predicting attitudes of interest. We assigned each one of 30 domain categories to subcategories: entertainment, consumption, general, media, life, consumption, and color-coded these categories in two palettes, orange and blue, to distinguish between categories that are related to life-style (gambling, shopping, entertainment etc.) and communication domains (messengers, social media etc.) The figure shows two visible patterns: top five domain categories that explain the most of the variance of attitudes towards democracy are related to media and communication, attitudes towards immigrants and populism can be best predicted with consumption and life style domains. Interestingly, entertainment has very low explanatory power, and issue oriented attitudes do not demonstrate any patterns in domain importance rank. Looking at specific attitudes, attitudes towards Islam can be explained best with media domains, and trust in national parliament — with communication related domains.

Domain level analysis with SVD. We also explore which websites specifically are associated with political attitudes which we do with Method II. For each political attitude with stable predictions and cross-validation, we took only SVD components with which they have significant correlation. We then filtered top 20 domains with positive loadings, and top 20 domains with negative loadings within each component, signifying the direction in which every domain is associated with the attitude of interest. Figure 4 illustrates association of political attitudes with domains, where SVD loadings are the x-axis, and domains color-coded into categories are on y-axis. Since from Method I, attitudes towards democracy and interest in politics demonstrated the best pre-

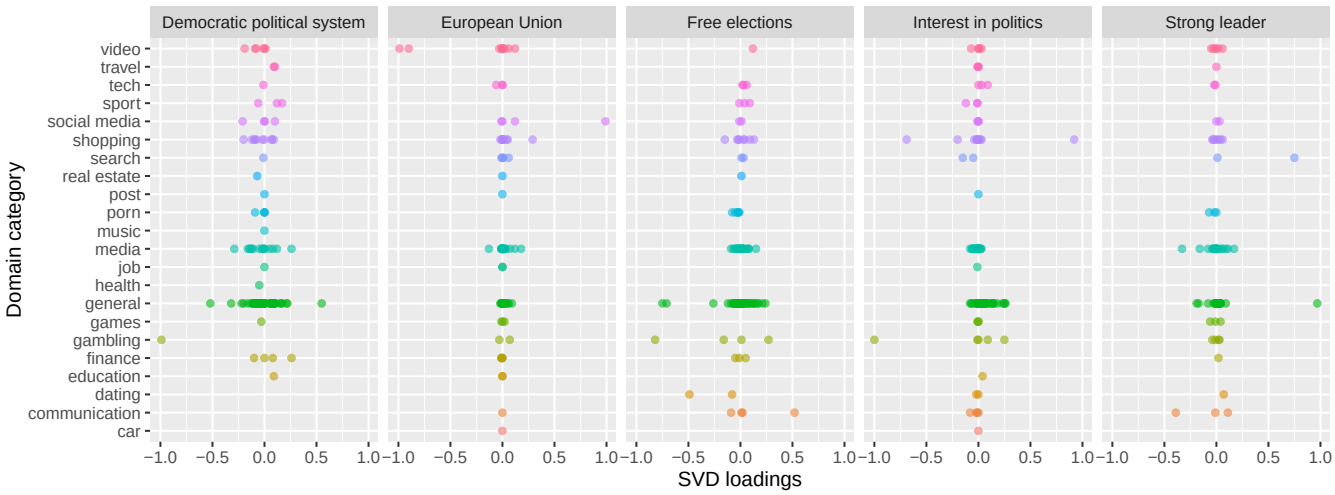

Figure 4: Association between attitudes towards democracy (and politics in general) and website domains. X-axis represents SVD loadings. Negative loadigns signify negative relation with an attitudes, and positive loadings — positive relation.

dictability, we focus on them in the analysis with SVD. Figure 4 summarise relationship between domains and attitudes towards democracy, European Union, and interest in politics. Although there are very general domain categories such as search, from which is hard to interpret, there are associations that can be telling in terms of social implications. Visits to social media, finance, communications (messengers), are associated with pro-democratic attitudes. One possible mechanism that could explain this connection is that people, who are monitor their finances, and communication with other people online might experience more fulfillment in their lives. Social media are also positively correlate with attitudes towards the EU, which is inconsistent with the existing research. Gambling, on the other hand, is strongly negatively associated with attitudes towards democratic political system, implying that people who are often loosing money are frustrated with democracy. Consistent with this findings, shopping is also on average negatively associated with attitudes towards democracy. The fact that video hosting platforms, streaming websites, and online gaming appear with both negative or positive or with very low loadings, implying that these websites plays significant role for people with either democratic or authoritarian attitudes, contrary to the public debate. Further in depth text-analysis of URLs is required to learn about differences in social media use between individuals with left and right ideological views.

Regression model with SVD components. We use SVD components which consist of domains with loadings, as predicting variables in a model. We fit generalized linear regression models using the k dimensions of SVD as independent variables for each of our dependent variables (survey items). We can write the regression models formally as follows:

$$Political\_Attitude = \alpha + \beta_1 K_1 + \beta_2 K_2 + ... + \beta_n K_n, \quad (2)$$

where $Political\_Attitude$ represents survey-based attitudes, α is the intercept, and β is a regression coefficient for every SVD component K. Overall, we have 16 questions measuring political attitudes, which we include into the model one by one.

On the project repository, we uploaded figures that report Pearson correlation coefficients by every SVD component k and averaged by the cross-validation folds for political attitudes with best predictions. The estimations show that web browsing behavioral patterns can predict individuals interest in politics, attitudes towards democratic political system, and Euroscepticism. The coefficients become much larger when we apply visit duration thresholds: at least one and five minutes. Interest in politics has median r = 0.09, attitude towards democratic political system — 0.07, and r = 0.04 for Euroscepticism, all obtained with k = 125. Best predictions are r = 0.15, 0.13, and 0.15 respectively, with k = 70, 80, and 60. The level of accuracy is compatible even with age, which has median r = 0.08 also with k = 125, and best prediction r = 0.22 with k = 50 (age usually serves as a benchmark in the literature, because it is a continuous variable, as opposed to a scale, and it is often reported accurately in surveys).

Conclusion

The results of this study demonstrate that information about website choices available from individuals' browsing histories can signal limited number of political attitudes. Our fine-grained dataset on 1,000 panelists in Germany, which generated approximately 19,000,000 URLs after three months of tracking, showed that interest in politics, attitudes towards democracy, perception of Islam, and trust in public institutions were especially identifiable from the web tracking data. Our straightforward estimation also shows that browsing histories are not informative to learn about attitudes towards climate change and populism, as well as other immigration related issues. However, in order to filter for relevant websites that meaningfully inform about

individuals' lifestyle and values, we applied a five-minute threshold to filter out accidental or not significant websites and the results improved drastically pointing out to the importance of weighting that every website has in the daily life of an individual.

Although we performed several tests of our sample and the data on generalizability, our results represent a conservative estimation of the predictive power of web browsing data. Our estimation is based on bounded ordinal independent variables that are common in the political science literature to measure political attitudes, but not always informative for machine learning models. With larger samples, better representations of URLs that are not limited to domains, alternative continuous instead of categorical measure of attitudes, and various model specifications, we expect the findings to gain more accuracy and robustness.

Despite the limitations of our data and measurement, the results are compatible with previous studies of individuals' personalities with larger samples. Our highest predictions for interest in politics and attitudes towards democracy vary from $r = 0.09$ to 0.15 compared to 0.17 for satisfaction with life also measured on five-point scale in Kosinski, Stillwell, and Graepel (2013), [0.20, 0.40] average estimation in Stachl et al. (2020) and in Funder and Ozer (2019). However, political attitudes related to populism and immigration demonstrated only relative stability in cross-validations, sample dependence, and low although still significant correlation coefficients, which is still consistent with the literature on classifying political orientation with social media data Cohen and Ruths (2013), and predicting life outcomes Salganik et al. (2020).

In summary, this paper introduces an empirical strategy for analyses of web tracking data with application to political behavior and attitudes. The study also makes suggestions for further research that seeks to explain political phenomena through the analysis of browsing histories. Finally, our results have important implications for policy makers in digital privacy and the society in general by emphasising beneficial as well as disadvantages for democracy potential of web browsing behavior to reveal people's political attitudes and issue preferences.

References

Agarwal, P.; Joglekar, S.; Papadopoulos, P.; Sastry, N.; and Kourtellis, N. 2020. Stop tracking me bro! differential tracking of user demographics on hyper-partisan websites. In *Proceedings of The Web Conference 2020*, WWW '20, 1479–1490. New York, NY, USA: Association for Computing Machinery.

Arcuri, L.; Castelli, L.; Galdi, S.; Zogmaister, C.; and Amadori, A. 2008. Predicting the vote: Implicit attitudes as predictors of the future behavior of decided and undecided voters. *Political Psychology* 29(3).

authors, A.

authors, A.

authors, A.

Bachrach, Y.; Kosinski, M.; Graepel, T.; Kohli, P.; and Stillwell, D. 2012. Personality and patterns of facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, 24–32. New York, NY, USA: Association for Computing Machinery.

Cerina, R., and Duch, R. 2020. Measuring public opinion via digital footprints. *International Journal of Forecasting*. To appear.

Cohen, R., and Ruths, D. 2013. Classifying political orientation on twitter: It's not easy! In *Seventh international AAAI conference on weblogs and social media*.

Dalton, R. J. 2000. Citizen attitudes and political behavior. *Comparative Political Studies* 33(6-7):912–940.

DellaPosta, D.; Shi, Y.; and Macy, M. 2015. Why do liberals drink lattes? *American Journal of Sociology* 120(5):1473–1511.

Englehardt, S., and Narayanan, A. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, 1388–1401. New York, NY, USA: Association for Computing Machinery.

Evans, D. C.; Gosling, S.; and Carroll, A. 2008. What elements of an online social networking profile predict target-rater agreement in personality impressions? In *ICWSM*.

Fatke, M. 2017. Personality traits and political ideology: A first global assessment. *Political Psychology* 38(5):881–899.

Funder, D. C., and Ozer, D. J. 2019. Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science* 2(2):156–168.

Garrett, R. K. 2009. Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate. *Journal of Communication* 59(4):676–699.

Gauchat, G. 2018. Trust in climate scientists. *Nature Climate Change* 8(6):458–459.

Gerber, A. S.; Huber, G. A.; Doherty, D.; and Dowling, C. M. 2010. Personality and political attitudes: Relationships across issue domains and political contexts. *American Political Science Review* 104(1).

Golbeck, J.; Robles, C.; and Turner, K. 2011. Predicting personality with social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, 253–262. New York, NY, USA: Association for Computing Machinery.

Golub, G. H., and Reinsch, C. Singular value decomposition and least squares solutions. 14(5):403–420.

Guess, A. M.; Nyhan, B.; and Reifler, J. 2020. Exposure to untrustworthy websites in the 2016 us election. *Nature Human Behaviour* 4(5):472–480.

Harteveld, E.; Kokkonen, A.; and Dahlberg, S. 2017. Adapting to party lines: the effect of party affiliation on attitudes to immigration. *West European Politics* 40(6).

Hu, J.; Zeng, H.-J.; Li, H.; Niu, C.; and Chen, Z. 2007. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, 151–160. New York, NY, USA: Association for Computing Machinery.

Kerna, M. L.; McCarthyb, P. X.; Chakrabartyb, D.; and Rizoiud, M.-A. 2019. Social media-predicted personality traits and values can help match people to their ideal jobs. *Proceedings of the National Academy of Sciences of the United States of America* 116(52).

Klimecki, O. M.; Vétois, M.; and Sander, D. 2020. The impact of empathy and perspective-taking instructions on proponents and opponents of immigration. *Humanities and Social Sciences Communications* 7(1):91.

Kosinski, M.; Stillwell, D.; Kohli, P.; ; Bachrach, Y.; and Graepel, T. 2012. Personality and website choice. In *ACM Web Sciences 2012*. ACM Conference on Web Sciences.

Kosinski, M.; Bachrach, Y.; Kohli, P.; Stillwell, D.; and Graepel, T. 2014. Manifestations of user personality in website choice and behaviour on online social networks. *Mach Learn* 95(3).

Kosinski, M.; Wang, Y.; Lakkaraju, H.; and Leskovec, J. 2016. Mining big data to extract patterns and predict real-life outcomes. 21(4):493–506.

Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America* 110(15).

Kramer, A. D. I.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America* 111(24).

Krosnick, J. A. 1991. The stability of political preferences: Comparisons of symbolic and nonsymbolic attitudes. *American Journal of Political Science* 35(3):547–576.

Kunda, Z. 1990. The case for motivated reasoning. *Psychological bulletin* 108(3):480—498.

Lambiotte, R., and Kosinski, M. 2014. Tracking the digital footprints of personality. *Proceedings of the IEEE* 102(12):1934–1939.

Lo Iacono, S. 2019. Law-breaking, fairness, and generalized trust: The mediating role of trust in institutions. *PLOS ONE* 14(8):1–14.

Prior, M. 2013. Media and political polarization. *Annual Review of Political Science* 16(1):101–127.

Rooduijn, M. 2018. What unites the voter bases of populist parties? Comparing the electorates of 15 populist parties. *European Political Science Review* 10(3):351–368.

Salganik, M. J.; Lundberg, I.; Kindel, A. T.; Ahearn, C. E.; Al-Ghoneim, K.; Almaatouq, A.; Altschul, D. M.; Brand, J. E.; Carnegie, N. B.; Compton, R. J.; Datta, D.; Davidson, T.; Filippova, A.; Gilroy, C.; Goode, B. J.; Jahani, E.; Kashyap, R.; Kirchner, A.; McKay, S.; Morgan, A. C.; Pentland, A.; Polimis, K.; Raes, L.; Rigobon, D. E.; Roberts, C. V.; Stanescu, D. M.; Suhara, Y.; Usmani, A.; Wang, E. H.; Adem, M.; Alhajri, A.; AlShebli, B.; Amin, R.; Amos, R. B.; Argyle, L. P.; Baer-Bositis, L.; Büchi, M.; Chung, B.-R.; Eggert, W.; Faletto, G.; Fan, Z.; Freese, J.; Gadgil, T.; Gagné, J.; Gao, Y.; Halpern-Manners, A.; Hashim, S. P.; Hausen, S.; He, G.; Higuera, K.; Hogan, B.; Horwitz, I. M.; Hummel, L. M.; Jain, N.; Jin, K.; Jurgens, D.; Kaminski, P.; Karapetyan, A.; Kim, E. H.; Leizman, B.; Liu, N.; Möser, M.; Mack, A. E.; Mahajan, M.; Mandell, N.; Marahrens, H.; Mercado-Garcia, D.; Mocz, V.; Mueller-Gastell, K.; Musse, A.; Niu, Q.; Nowak, W.; Omidvar, H.; Or, A.; Ouyang, K.; Pinto, K. M.; Porter, E.; Porter, K. E.; Qian, C.; Rauf, T.; Sargsyan, A.; Schaffner, T.; Schnabel, L.; Schonfeld, B.; Sender, B.; Tang, J. D.; Tsurkov, E.; van Loon, A.; Varol, O.; Wang, X.; Wang, Z.; Wang, J.; Wang, F.; Weissman, S.; Whitaker, K.; Wolters, M. K.; Woon, W. L.; Wu, J.; Wu, C.; Yang, K.; Yin, J.; Zhao, B.; Zhu, C.; Brooks-Gunn, J.; Engelhardt, B. E.; Hardt, M.; Knox, D.; Levy, K.; Narayanan, A.; Stewart, B. M.; Watts, D. J.; and McLanahan, S. 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences* 117(15):8398–8403.

Settanni, M.; Azucar, D.; and Marengo, D. 2018. Predicting individual characteristics from digital traces on social media: A meta-analysis. *Cyberpsychology, Behavior, and Social Networking* 21(4):217–228. PMID: 29624439.

Shi, F.; Shi, Y.; Dokshin, F. A.; Evans, J. A.; and Macy, M. W. 2017. Millions of online book co-purchases reveal partisan differences in the consumption of science. *Nature Human Behaviour* 1(4).

Stachl, C.; Au, Q.; Schoedel, R.; Gosling, S. D.; Harari, G. M.; Buschek, D.; Völkel, S. T.; Schuwerk, T.; Oldemeier, M.; Ullmann, T.; Hussmann, H.; Bischl, B.; and Bühner, M. 2020. Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences* 117(30):17680–17687.

Wiedemann, A. 2020. Austerity, indebtedness, and political behavior. evidence from the u.k. *Working paper*.