Yasseri and Bright *EPJ Data Science* (2016) 5:22
DOI 10.1140/epjds/s13688-016-0083-3

EPJ.org

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Wikipedia traffic data and electoral prediction: towards theoretically informed models

Taha Yasseri[*†] and Jonathan Bright[†]

[*]Correspondence:
taha.yasseri@oii.ox.ac.uk
Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford, OX1 3JS, UK
[†]Equal contributors

Abstract

This aim of this article is to explore the potential use of Wikipedia page view data for predicting electoral results. Responding to previous critiques of work using socially generated data to predict elections, which have argued that these predictions take place without any understanding of the mechanism which enables them, we first develop a theoretical model which highlights why people might seek information online at election time, and how this activity might relate to overall electoral outcomes, focussing especially on information seeking incentives related to swing voters and new parties. We test this model on a novel dataset drawn from a variety of countries in the 2009 and 2014 European Parliament elections. We show that while Wikipedia offers little insight into absolute vote outcomes, it does offer good information about changes in overall turnout at elections and about changes in vote share for particular parties. These results are used to enhance existing theories about the drivers of aggregate patterns in online information seeking, by suggesting that voters are cognitive misers who seek information only when considering changing their vote.

Keywords: social data; elections; prediction; big data; Wikipedia; public opinion

1 Introduction

As digital technologies become more and more integrated into the fabric of social life their ability to generate large amounts of information about the opinions and activities of the population increases. The potential applications of what could be described as 'socially generated data' (which range from mobile phone GPS records to content produced on Twitter and searches made on Google) are widespread in terms of new understanding of human behaviour [1, 2]. A particular focus of research has been on the possibility of using these data for predicting a wide range of social phenomena [3] such as stock market fluctuations [4, 5], outcomes in public health [6, 7], movie box office revenues [8], unemployment [9], and election results [10–13]. The opportunities in this area are enormous: predictions based on socially generated data are much cheaper than conventional opinion polling, offer the potential to avoid classic biases inherent in asking people to report their opinions and behaviour, can deliver results much quicker and be updated more rapidly, and can offer purchase on phenomena (such as stock market movements) for which there are no well-established forecasting models.

However, whilst a variety of positive results have been published, some strong criticisms of socially generated predictions have also been raised, particularly in the field of electoral prediction (which, perhaps because of the widespread availability of validation data or because of its importance for conventional opinion polling has been the most frequent application of this type of prediction). Several authors have argued that work which has shown correlations between socially generated data and electoral outcomes has been done 'without an analysis of what principle enables them' [14]. For example, given what is known about how Twitter's user base differs in demographic characteristics from the general population [15, 16], and also the fact that people using Twitter do not necessarily have an incentive to tweet about their voting intention, there is little theoretical reason to expect that the overall volume of tweets about different politicians will be proportionate to their overall volume of votes. This has led several authors to argue that the positive correlations which have been observed so far between Twitter and electoral outcomes (or indeed any socially generated data) might simply have been achieved by chance [17], with Metaxas *et al.* arguing that 'predicting elections with accuracy should not be without some clear understanding of why it works' [18].

Our approach in this article is inspired by the challenge laid down by some of these critiques. The aim is to develop a theoretically informed prediction of election results from socially generated data, which is based not just on observation of correlation between raw numbers and eventual outcomes but also an understanding of the social processes through which the numbers are generated. We also aim to apply the models developed to a variety of different countries in two separate elections, thus offering a more general test of their usefulness than work which has focussed on just one country. Through this process we hope to both explore the predictive power of socially generated data and enhance theory about the relationship between socially generated data and real world outcomes. Our particular focus is on the readership statistics of politically relevant Wikipedia articles (such as those of individual political parties) in the time period just before an election. Wikipedia is one of the most popular sources of information online, with recent survey evidence finding that large proportions of the adult population make use of it in a variety of different country contexts [19, 20]. This makes it a good candidate for potential use for predictive purposes.

We will begin by outlining a theory of the relationship between Wikipedia page view statistics and overall electoral outcomes. We base this theory on a rational choice approach to explaining voting behaviour [21], which conceptualises voters as similar to consumers in a market, seeking to vote for the political party who offers them the greatest 'pay-off' in terms of policies.[a] Online information seeking, from this rational choice perspective, can be explained as the result of voters looking for more information about the election: perhaps about practical matters such as how to vote, or perhaps about substantive matters such as which political party might suit them best. Such information seeking is rational in that it increases the chance that a voter can pick a party which represents them well and thus improve their pay-off from the election. Hence we assume that page views to political pages in the period before an election are generated by voters seeking information about the election. Of course, Wikipedia page view traffic will also include journalists writing stories about the parties, 'opinion leaders' looking for information to pass on to their social connections [22], and those working for the parties themselves checking to see pages are up to date, however we assume the number of these people is small in relative terms, and furthermore relatively evenly distributed across all parties, hence we do not include them in the theory.

If this theory is correct, increases in Wikipedia page views ought to have predictive value. In particular, increases in Wikipedia views to politically relevant pages ought to indicate increases in turnout at election time, both in terms of overall turnout of voters at the election and in terms of turnout to vote for specific parties, as they should indicate that more voters are considering voting in the election for a given party (this idea is supported by research that has demonstrated that those looking for political information online are also more likely to participate in general [23]). Of course, we do not know if the person looking for information on Wikipedia will actually turn out and vote. We also do not know if they have been convinced by the Wikipedia page of a political party to vote for that party. But at the aggregate level an increase in people considering a vote ought to correlate with an increase in actual votes. This logic is analogous to the logic behind the use of Wikipedia page views to forecast consumer behaviour in other areas such as stock market movements [4] and movie box office revenues [8]: while the authors of these papers do not claim, for example, that every person reading the Wikipedia page about a film automatically goes to see it, larger volumes of readership ought to indicate increased interest in that film which should translate into higher box office receipts.

However, we also expect that a number of factors will moderate the relationship between online information seeking and electoral outcomes, meaning that Wikipedia page views may not be much use as a direct predictor unless certain corrections are included. Rational voters are 'cognitive misers' [24] who will not seek new information unless they think it necessary, thus minimizing the costs of voting. This has a number of potential consequences. People are more likely to seek information on new political parties as they are less likely to have a pre-established opinion on whether this party offers them a good payoff (an idea supported by research which has indicated that online searches on Google are more likely for unfamiliar political topics [21]). They are also likely to consider the perceived 'viability' of a party (*i.e.*, whether it is likely to be able to contend as a serious electoral force): people are less likely to seek information on parties which do not appear to have a chance of seriously contending [25] (this may also serve to create a kind of 'spiral of silence' around minor parties who, because they are perceived to be less viable, are systematically ignored when people seek information [26]).

In addition to being more likely to seek information on new parties, people are also more likely to seek information if they are considering changing their vote (if they are sticking with a party they have chosen previously, the need for new information is lower). This might emerge from dissatisfaction with the party for whom they voted at the previous election (for example, studies have shown that dissatisfaction with government can promote higher online information seeking [27]). Hence Wikipedia page views may be driven more by 'swing voters' who are switching to a new party, and who wish to inform themselves about their choices, rather than voters who are voting for the same one again. Recent studies have shown support for the idea that swing voting is associated with information seeking by showing that these voters typically have at least some amount of political knowledge [28], and that very well informed 'apartisans' (who are always considering changing their vote) now constitute an important part of the electorate [29].

In addition to the relationship between Wikipedia page views and electoral outcomes, any theory of online information seeking and politics also needs to take into account the potential influence of the news media, which has traditionally been the venue where people seek politically relevant information [30]. Information patterns within the news media may themselves correlate with electoral outcomes to a significant extent, as media outlets will often seek to make their coverage more or less proportional to the importance of different political parties. However, coverage of older parties is perhaps likely to be more widespread, especially if they are incumbent in the current government. News media coverage may also have an impact on Wikipedia page views, in both a positive and negative sense: being mentioned in the news media might stimulate people to find out more about a candidate online; but it might also fulfil people's information needs, and hence make them less likely to seek information from alternative sources.

This theory of how online information seeking relates to eventual outcomes suggests that a model which uses Wikipedia page views to predict electoral outcomes should take into account several factors. First, we would expect new parties, already established parties and incumbent parties to experience different page view dynamics, with newer parties receiving a disproportionately large amount of page views compared to their final number of votes (a result already observed in socially generated predictions based on Twitter data [16]), whilst more established parties experience the reverse effect. Second, parties attracting lots of swing voters may also do disproportionately well in page view statistics, whilst parties which are losing votes may again do disproportionately badly. Finally, coverage of the political party in the mainstream news media may well correlate with voting behaviour itself, as well as serving to 'replace' any effect of Wikipedia page views (hence parties which are well covered by the news media are may be badly covered by Wikipedia).

2 Data

Our aim in this paper is to test the extent to which models can be developed using Wikipedia data to predict electoral outcomes (both in terms of turnout and especially in terms of results for individual parties), and also the extent to which these models can be improved using these simple theoretically informed corrections. In order to do this, we built a dataset centred on the two most recent European Parliament elections (in 2009 and 2014). The European Parliament elections were chosen because they would allow us to build a relatively large sample of political parties, all competing at the same time under broadly the same electoral system. However, it should be noted that this focus is somewhat of a limitation as European elections are typically perceived as secondary elections in most EU member states, subordinate to the national electoral contest.

We collected two types of data on these elections. First, page view statistics for the general Wikipedia page on the election were harvested in 14 different language editions of Wikipedia, each one representing one of the countries which went to the polls on election day.[b] These 14 language editions were chosen based on the following criteria: (a) they are the primary spoken language of one country, (b) the country that the language is spoken in has been an EU member in the last two rounds of elections, and (c) the corresponding Wikipedia pages existed prior to the election date in that language edition. They were also chosen because they have relatively strong user bases in Wikipedia, which provides a good basis for comparison.

Second, we created a dataset of political parties which competed in either or both of the 2009 and 2014 elections in the five largest Western European countries: the UK, France, Germany, Spain and Italy. Parties which competed in both are represented twice, with one observation for each election. The full list of parties is available in Additional file 1. We chose to focus only on those parties which secured more than 5% of the vote in either year (even if a party scored 5% in 2009 and 4% in 2014, it is still included for both elections). European Parliament elections (like any election) typically feature a small number of parties which absorb the vast majority of votes, and a long tail of more minor parties which achieve little or no electoral success [31]. Having such a threshold is therefore a necessary simplification, as it removes a long tail of very minor parties who otherwise would have constituted the majority of the observations. In total the dataset contains 59 observations.

For each party, we recorded several variables of interest. First, we recorded the amount of views the page of the political party in question in the corresponding Wikipedia language edition received in the week before the election. There was some ambiguity about which Wikipedia page to use in some cases, especially where more than one party presented itself in the form of a coalition, and hence more than one page could be valid. In these cases, we always used the Wikipedia page which had the highest number of views. A full list of the Wikipedia pages used is available in Additional file 1. Results of the 2009 and 2014 elections were taken from the official web page of the elections.[c] Results of the 2004 elections, which were needed to calculate the change between 2004 and 2009, were taken from the Norwegian Centre for Research Data.[d] We then also recorded the percentage of the vote achieved by the party, and the difference between that percentage and the previous year, in order to try and captures whether voters swung towards or away from the party. Change in vote share is, of course, not identical to the amount of swing voters: a party could, for example, gain and lose the same number of votes, which would mean lots of swing voters but no visible aggregate change. However, we expect the amount of swing voters and the amount of change to be correlated.

We also recorded several variables which allow us to apply our theoretically informed corrections. First, we recorded whether each political party was 'new' or not at the time of the election. Newness is a somewhat ambiguous category, as many of the parties which surged to prominence in the 2014 elections had existed for a long time as relatively minor political forces. Many apparently new parties are also simply existing ones which have been rebranded with a new name. Our aim with this variable was to capture the extent to which the majority of the electorate was likely to recognise this party already. Hence parties were listed as new not only if they did not compete at the previous election, but also if they either had a different name to that which they competed under in the previous election, or scored less than 5% at the previous election. We made two exceptions to this rule: the major centre left and centre right parties in Italy both changed their names before the 2009 election; however, we did not record them as 'new' parties because they were the major political forces in Italy at the time. Incumbency was a simpler variable to measure: it simply records whether the party was an incumbent in the national government at the time of the election, which would again likely change the extent to which they are visible to the electorate. Incumbency and newness are related of course, as a party cannot be both new and incumbent: it can however be neither new nor incumbent, hence it is useful to have both variables.

Finally, we recorded the amount of times the party was mentioned in the print media in the week immediately prior to the election. These numbers were calculated by conducting a search for the party in the LexisNexis news media dataset[e] in the largest official language of the country in question. This dataset is a large archive of material produced by print newspapers all over the world, which can act as a useful proxy of media attention to an individual party. The number of search results returned was used as the number of media 'mentions' recorded (with one mention being one news article which contained at least one reference to the party in question). This variable is of course not a perfect measure. In particular, in many cases the exact search term to use in LexisNexis was not self-evident. For example, the Christlich Demokratische Union in Germany is often referred to as the CDU. In cases where more than one search term was potentially valid, we conducted the search with all possible terms, and used the highest number which resulted. We did however avoid terms which are also commonly used words, for example 'greens' as a shorthand for the Green Party. A full list of search terms used is available in the Additional file 1.

3 Results

Our results section is divided into three parts. First, we look at the relationship between Wikipedia traffic patterns around election time and overall electoral turnout. Then, we develop a model to try and predict absolute vote share outcomes for different political parties. Finally, we look at whether such a model can be developed for changes in vote share.

We will begin by looking at the relationship between overall levels of traffic to the general political articles in different language editions of Wikipedia and electoral turnout. These articles offer general information about the election, such as the date on which it will be held. As we highlight in our theory section we expect that more people looking at this page indicates that higher levels of people are interested in voting in the election, and are seeking to inform themselves about the practicalities of voting.

Figure 1(a) shows the overall pattern of page views of the main Wikipedia article in the EU Parliament elections around the election time in 2009 in 14 different language editions of Wikipedia. Despite all the differences in the political settings of the associated countries, some common patterns are evident from this diagram: a gradual increase can be observed, starting about a month before the election date, followed by a large jump a few days before the election and finally a decay after the results are announced. It seems that the vast majority of information seeking about elections takes place in the few days before the election itself. These data further justify a focus on page views in the week just before the

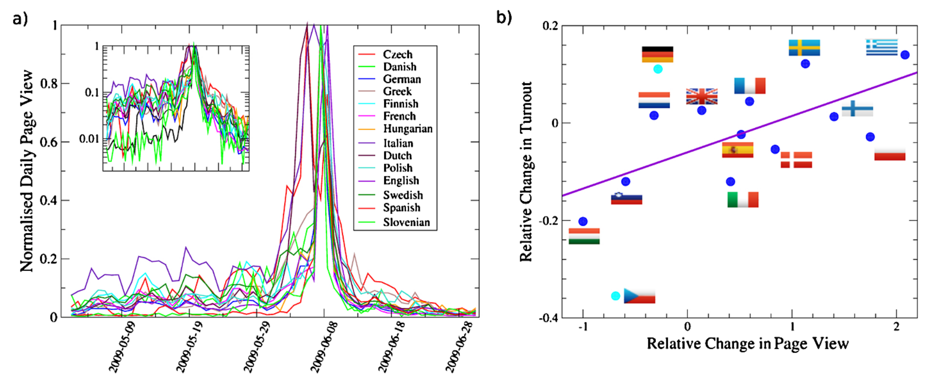

Figure 1 EU elections and Wikipedia page views. (a) Page views of the general Wikipedia article on the European Parliament elections over time and **(b)** relative change in these page views compared to change relative change in turnout. The two outliers, Germany and Czech Republic are excluded from the trend line in (b).

election. From the inset of Figure 1(a), which shows the same data in logarithmic scale, we observe that the absolute value of the decay rate in the post-event attention is larger than the one of the pre-election build up phase (similar to observations in [24]). In other words, interest in the event fades very quickly. The peaks of attention fall on different days: we attribute this to the fact that the EU election falls on different days (and sometimes spans multiple days) in different countries.

If our theory about the reasons for online information seeking holds true, then the volume of attention in the build-up phase before the election should be an indicator of the general level of interest in that election (in particular, whether people are considering a vote at all), and therefore a predictor for the overall turnout in each country. Of course, different language editions of Wikipedia have very different sized user bases, meaning that the absolute level of attention to the general Wikipedia article is not likely to be of much use as a predictor. Hence we look instead at the relative change in page views to this general Wikipedia article between the 2009 and 2014 European elections, and compare this to the relative change in turnout between those two elections. Figure 1(b) plots the correlation between these two values. The strength of the correlation ($R = 0.59$, adjusted $R^2 = 0.29$, p-value = 0.04) is reasonable considering both the limitations of the data and the simplicity of the model. If we remove two outlying points (the Czech Republic and Germany) the correlation of $R = 0.72$ (adjusted $R^2 = 0.47$, p-value = 0.004) improves considerably. This shows some good initial support for our theory that general levels of interest in a political event are proportional to general levels of readership on Wikipedia. However, we do not have a good explanation for the different dynamics which may have produced the two outliers (the Czech Republic in particular does not seem to fit the overall pattern).

We now move on to our main focus, which is predicting the performance of individual parties in the two elections and five countries as described above. We focus again on page views in the week before the election, this time to the Wikipedia page of the individual party in question in the language edition relevant to that country. Again, as language editions of Wikipedia have different volumes of traffic, we conduct another normalisation, but this time to the sum of page views of all parties competing in the same country at the same time. This can be thought of, in other words, as a party's 'share' of Wikipedia traffic for that election.

As we describe above, theoretically we do not expect much of a correlation between absolute levels of Wikipedia traffic and absolute levels of vote share, and indeed in practice we observe $R = 0.05$. We explore the reasons for this in Figure 2, which shows the relative levels of traffic each party achieved, separated for each country/election. We observe that the 'winner' in terms of Wikipedia vote share is, of course, not the winner in terms of overall vote share, but is instead often a relatively small 'anti-establishment' party who ended up doing surprisingly well. This is the case of, for example, the Front Nationale (FN) in France in 2014, the British National Party (BNP) in the UK in 2009 and Podemos in Spain in 2014. This fits in with some of the theory we described above: these parties were relatively minor political forces at the time, and some were completely new, but they attracted a lot of swing voters in those elections. Hence they were doubly likely to be favoured in page view statistics: lots of voters were considering voting for them, and many of these voters would have had little prior information about these parties.

As we highlight in our theoretical discussion, we expect several corrections to be necessary to Wikipedia view data in order for it to act as an effective predictor. We will exam-

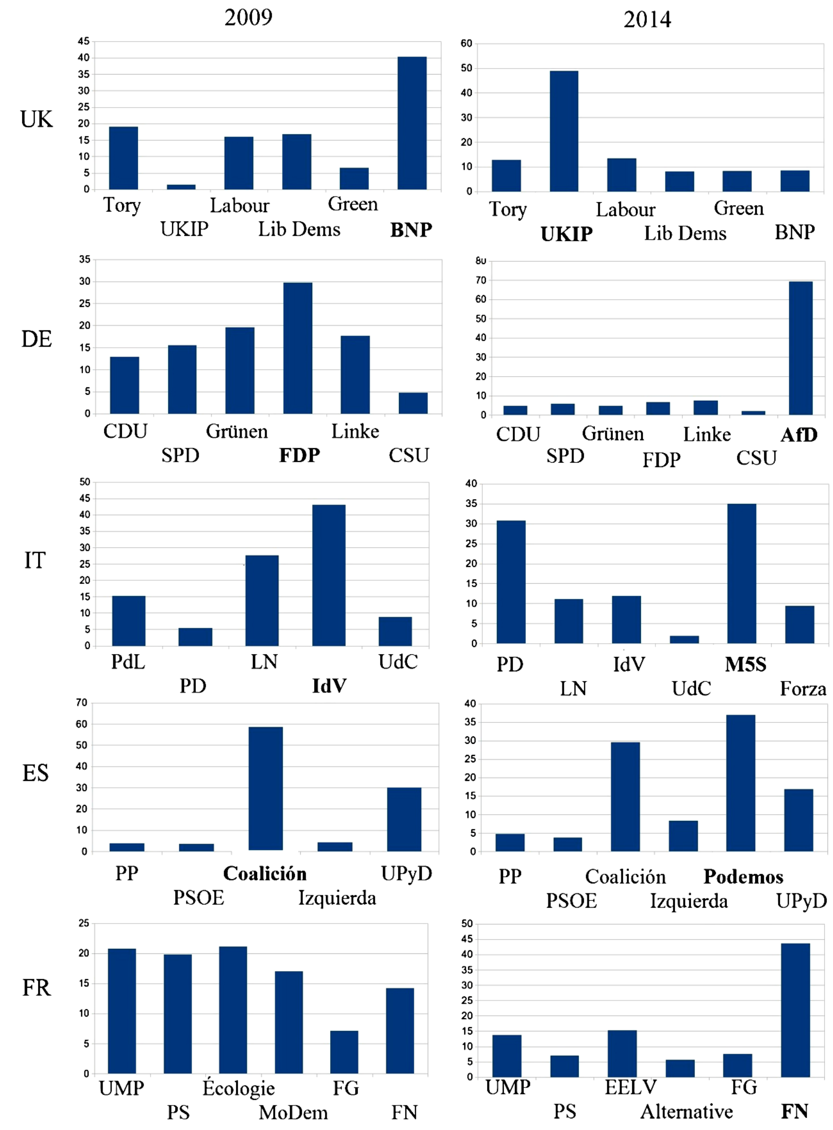

Figure 2 Wikipedia page view statistics for different parties. Relative share of Wikipedia traffic for major parties in five European countries the week before the 2009 and 2014 European parliament elections (for the party names, see the list of abbreviations below).

ine first of all the news media. Figure 3(a) shows the correlation between news mention counts of political parties in the week before the election and the overall vote share. The correlation is actually very high ($R = 0.84$, adjusted $R^2 = 0.71$, p-value < 0.00001). As we described above, news outlets tend to calibrate their coverage to the perceived importance of political parties, hence this correlation is not surprising. However, it is notable that it is also much stronger for the larger parties, whilst only moderate for the smaller parties shown in the shaded area of the graph ($R = 0.52$, adjusted $R^2 = 0.25$, p-value $= 0.0005$).

We have observed that Wikipedia seems to predict the successful emergence of new parties very well and news media is a good indicator for the more established parties, we can further deepen this observation by comparing Wikipedia attention to news mentions. We suggested above that the news media might drive traffic to Wikipedia (which would make

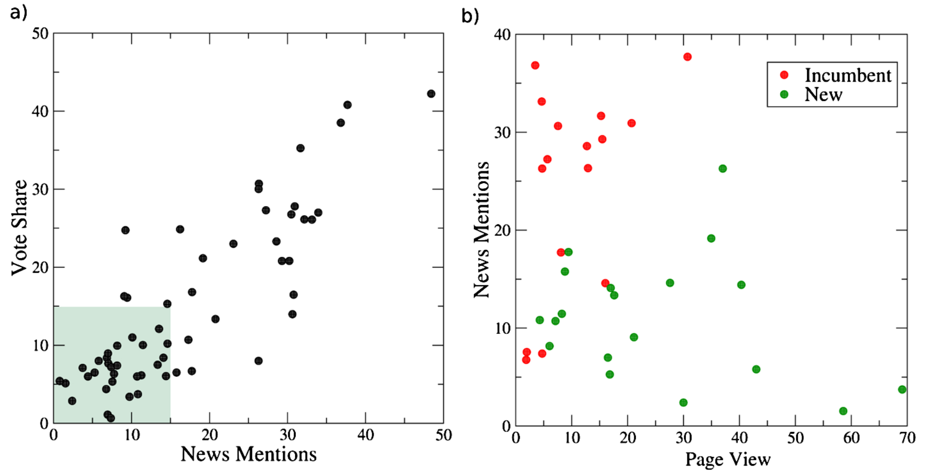

Figure 3 Attention and media effect. News media mentions compared to **(a)** Wikipedia page views and **(b)** absolute level of vote share.

Wikipedia page views essentially epiphenomenal), or that it might also 'replace' Wikipedia (with more news mentions leading to less views). However, as we see in Figure 3(b), neither of these correlations can be observed. In Figure 3(b), parties are colour coded based on their political history: green for new parties and red for incumbent parties (with parties which were neither new or incumbent removed). We can see that there are two distinct clusters on this graph: incumbent parties which are over-represented in the media and under-represented in Wikipedia, and vice versa for new parties. This would suggest that news media mentions should be corrected to account for incumbency, whilst Wikipedia mentions are corrected to account for the newness of a party.

With these observations in hand, we will now move on to a more systematic investigation of the usefulness of Wikipedia page view data in predicting electoral outcomes. We divide this investigation into two parts. First, we look at the use of page view data in predicting absolute vote share outcomes for individual parties. Second, we look at the use of this type of data in predicting changes in vote share outcomes when compared to previous elections.

We will begin by looking at absolute vote share outcomes. As we highlighted above, we theorise that increases in Wikipedia page view counts may indicate increasing numbers of votes for a party, as voters who are considering a vote for that party may be informing themselves about its policies; hence Wikipedia data may be of use for predicting electoral results. We test this contention with three analytical models reported in Table 1. These models are linear OLS regressions looking at the relationship between voting outcomes (measured as a percentage vote share from 0-100) and predictors related to information seeking behaviour, and were estimated in *R*.[f] The first model, 1.0, is a baseline model, which looks simply at the relationship between the vote share outcome for a party in a European election and its result in the most recent national election. This is an obvious, freely available statistic which we would expected to provide a reasonable indication of future electoral performance. A baseline model making use of this statistic allows us to explore the extent to which Wikipedia parameters add useful new information when making predictions. Comparing a model based on Wikipedia data to this model therefore allows for quite a strong test of its usefulness.

Table 1 Predicting absolute vote share outcomes

	Model 1.0: Baseline		**Model 1.1: Baseline with corrections**		**Model 1.2: Full model**	
	β	**SE**	***β***	**SE**	***β***	**SE**
Intercept	5.80***	(1.16)	3.32	(1.68)	0.82	(1.94)
Previous national result	0.61***	(0.06)	0.29**	(0.09)	0.33***	(0.09)
News			0.45***	(0.10)	0.41***	(0.10)
New party			–0.73	(1.64)	1.26	(2.53)
Incumbency			–6.75	(3.57)	–4.91	(3.55)
News x incumbency			0.26	(0.15)	0.19	(0.15)
Wikipedia					0.18*	(0.08)
New party x Wikipedia					–0.15	(0.10)
R^2	0.67		0.79		0.81	
Adjusted R^2	0.66		0.77		0.78	
AIC	386.32		368.44		366.43	
BIC	392.55		382.98		385.13	
n	59		59		59	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

The second model, 1.1, is another baseline model which, in addition to the past voting results, contains all theoretically relevant parameters except those relating to Wikipedia. In particular, it contains parameters for the relative share in the news items (measured from 0-100), a dichotomous indicator for whether the party is a new party, another for whether it is an incumbent party (and we should highlight again that these two variables do not measure the same thing, as a party can be neither new nor incumbent), and an interaction effect between the news count and the incumbency indicator. As we discussed above, these terms are included in our full model primarily as a means of correcting the Wikipedia variable. However, presenting them in a model without the Wikipedia terms themselves allows us to specify the precise impact of Wikipedia itself on the predictive power of the model. Models 1.0 and 1.1 serve as a basis for comparison for model 1.2, which is a full model containing an additional two Wikipedia relevant parameters: a main effect for the relative share in Wikipedia page views (again measured from 0-100), and an interaction effect between Wikipedia views and status as a new party.

These models provide evidence for the idea that Wikipedia page view data is correlated with vote share outcomes; however, the improvement in predictive accuracy is very modest. We focus on adjusted R^2 as a way of measuring this accuracy as it takes into account not only the correlation between dependent and independent variables but also penalizes models for having more terms which do not necessarily include more information. It is thus a strong test of whether the addition of parameters increases the overall quality of the model (we also consider the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as further ways of testing this). The overall R^2 of these models is comparatively high. However, when compared to model 1.1, model 1.2 increases R^2 by just 0.02, which translates to an increase in adjusted R^2 of 0.01. While AIC is also reduced, the BIC of model 1.2 is actually increased with respect to model 1.1 (albeit fractionally). Hence overall while Wikipedia might technically be considered a predictor of absolute vote share outcomes, it can offer only a marginal improvement on baseline models.

As we are interested in enhancing the theory of online information seeking, it is also worth discussing the coefficients of individual indicators (here we focus on the results reported in model 1.2). As we would expect based on Figure 3(a), the news coefficient is positive, statistically significant in all models, and of a considerable size: for example, gaining 50% of the news mentions would, on average, correlate with an increase in vote share of around 20 percentage points. The term for Wikipedia also points in the expected direction is statistically significant: more Wikipedia page views do seem to correlate with more votes (50% of the Wikipedia 'page view share' would correlate with around 9 percentage points more votes). As we also expected, the new party interaction reduces the effect of the Wikipedia term considerably (though this term is not statistically significant). Hence it does indeed seem to be the case that Wikipedia page views overstate the potential electoral impact of new parties.

In terms of our theoretical discussion, it is of course dangerous to draw conclusions about micro-level behaviour on the basis of aggregate data. However, the results are consistent with the idea that page views on Wikipedia are driven by the rational model of information seeking we describe above: that is, people may be seeking information when they are considering a vote, but disproportionately higher levels of information seeking occur when parties are less well known. But not all of our theoretical suggestions were supported: the results undermine the 'viability' thesis (that is, the idea that new parties will attract a disproportionately small amount of information seeking as people do not perceive them as a viable option and hence do not bother considering to vote for them).

As we highlighted above, we expected Wikipedia to offer predictive power in the case of swing voters rather than in the case of absolute vote share: and we already have descriptive evidence that this is the case from Figure 2. Hence we will now move on to a second set of analytical models which try and address this question, which are presented in Table 2. Instead of trying to predict absolute amounts of vote share, these models try and predict change, that is the amount by which a party's vote share increases or decreases compared to the last election (with new parties considered to have been at 0 votes in the last election). This variable is measured from –100 to 100. As well as a new dependent variable, these models also employ the change witnessed at the previous national election

Table 2 Predicting change in vote share outcomes

	Model 2.0: Baseline		**Model 2.1: Baseline with corrections**		**Model 2.2: Full model**		**Model 2.3: Baseline with Wikipedia**	
	β	**SE**	β	**SE**	β	**SE**	β	**SE**
Intercept	–0.02	(0.89)	–0.43	(2.26)	–5.75*	(2.35)	–5.67***	(1.47)
Change in previous national result	0.46**	(0.14)	0.45**	(0.14)	0.37**	(0.13)	0.38**	(0.12)
News			–0.01	(0.11)	–0.02	(0.10)		
New party			2.52	(2.17)	5.08	(2.98)	4.78	(2.75)
Incumbency			–3.30	(4.81)	0.81	(4.33)		
News x incumbency			0.10	(0.20)	0.00	(0.18)		
Wikipedia					0.37***	(0.09)	0.36***	(0.09)
New party x Wikipedia					–0.26*	(0.12)	–0.24*	(0.11)
R^2	0.17		0.21		0.42		0.41	
Adjusted R^2	0.15		0.13		0.34		0.37	
AIC	398.69		403.81		389.9		384.11	
BIC	404.92		418.36		408.6		396.58	
n	59		59		59		59	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

as a baseline predictor (*i.e.*, the difference between the vote share obtained in the previous national election and the one before that), in order to provide a strong test of Wikipedia's usefulness. Apart from these differences, models 2.0-2.2 are identical to models 1.0-1.2.[g]

These models provide much better evidence for the potential usefulness of Wikipedia page views. The full baseline model (2.1) performs relatively poorly, with an adjusted R^2 of just 0.13 (and indeed the model is worse than model 2.0, which only uses the information about outcomes in national elections). However, the full model, 2.2, performs much better, with an adjusted R^2 of 0.34 and much lower AIC and BIC. Hence when it comes to predicting relative change in vote share, Wikipedia seems to provide useful information. The BIC of model 2.2 is actually higher than model 2.0, which contains only the national election indicator. However, this increase is related to the inclusion of parameters which were brought in model 2.1 (especially those relating to news and incumbency). To demonstrate this, we specify a final model, 2.3, which includes only the national vote change indicator and the terms relevant to Wikipedia. This model has the highest adjusted R^2 and lowest AIC and BIC of all of the second set of models.

In terms of individual results, the coefficient for the Wikipedia term is statistically significant, with parties which hold 50% of the Wikipedia share for a given country picking up a 17.5 percentage point increase in votes. This provides further support for the thesis that Wikipedia page views might be driven by people switching their vote from one party to another, and hence a strong surge in page views might indicate an increase in vote share for the party in question. The new party interaction term is also statistically significant, and reduces this effect considerably providing further evidence that Wikipedia page view statistics are disproportionately biased towards new parties. These results reinforce the theoretical remarks made about the first set of models.

4 Discussion and conclusions

In this paper we have sought to develop theoretically informed methods for election prediction based on information seeking behaviour on Wikipedia, responding to existing critiques of predictions generated from new sources of socially generated data. We applied these methods to a variety of different European countries in the context of two different European elections. We have produced three main empirical findings. First, we have shown that the relative change in the number of page views to the general Wikipedia page on the election can offer a reasonable estimation of the relative change in turnout for that election at the country level. This supports the idea that increases in online information seeking at election time are driven by voters who are considering voting in the election. Second, we have shown that a theoretically informed model based on previous national results, Wikipedia page views, news media mentions, and basic information about the political party in question can offer a good prediction of the overall vote share of the party in question. However, the Wikipedia variable itself was of relatively minor importance in this prediction. Third, we presented a model for predicting change in vote share (*i.e.*, voters swinging towards and away from a party). We showed that Wikipedia page view data provided for an important increase in predictive power in this context. We also showed, however, that this relationship is exaggerated in the case of newer parties.

Based on this information, we also enhanced our theory of the sources of Wikipedia traffic before election time. We have shown good evidence that newer parties which attracted a lot of swing voters received disproportionately high levels of Wikipedia traffic.

Our data is at an aggregate level therefore we cannot offer firm conclusions about individual behaviour: however, the data is consistent with the idea that voters do not seek information uniformly about all parties at election time. Rather, they behave like 'cognitive misers', being more likely to seek information on new political parties with which they do not have previous experience and being more likely to seek information only when they are actually changing the way they vote. By contrast, there was no evidence of a 'media effect': there was little correlation between news media mentions and overall Wikipedia traffic patterns. Indeed, the news media and Wikipedia appeared biased towards different things: with news favouring incumbent parties, whilst Wikipedia favoured new ones. Furthermore, there was no evidence of a 'viability' effect: voters did not appear to ignore parties that had little realistic chance of gaining power.

It is worth concluding by reflecting on the limitations of the work. We focussed on EU level elections: national elections may experience different dynamics, and would also be worth studying. Our measure of swing voters (change in vote share when compared to the previous year) is also imperfect, as voters may of course swing both towards and away from a party in an election. We focussed on relatively major political parties, which may explain our negative findings in terms of the 'viability' thesis: all of our parties were viable, to some extent. Furthermore, we do not know much about the content of the news article we measured: whether it was short or long, positive or negative. Such data would undoubtedly improve our understanding of the informational landscape surrounding parties at election time. Likewise, we do not know to what extent Wikipedia page views really are driven by voters, as opposed to other political actors such as journalists. Perhaps most importantly, our theories concern micro level behaviour, but our data is measured at an aggregate level, which may mask micro level effects. Future work could usefully address these concerns.

Additional material

Additional file 1: Party List: A table containing countries, name of the parties (English and local), election dates, party abbreviations, election vote share, change in the vote share from the previous election, number of news mentions, and the link to the Wikipedia page. (csv)

Abbreviations

AfD: Alternative für Deutschland; BNP: British National Party; Grünen: Bündnis 90/Die Grünen; CDU: Christlich Demokratische Union; CSU: Christlich-Soziale Union; Coalición: Coalición por Europa; Tory: Conservative Party; Linke: Die Linke; Écologie: Europe Écologie; EELV: Europe Écologie Les Verts; Forza: Forza Italia; FDP: Freie Demokratische Partei; FG: Front de gauche; FN: Front National; Greens: Green Party; PdL: Il Popolo della Libertà; IdV: Italia dei Valori; Izquierda: La Izquierda; Labour: Labour Party; LN: Lega Nord; Lib Dems: Liberal Democrats; MoDem: Mouvement Démocrate; M5S: Movimento 5 Stelle; PS: Parti Socialiste; PP: Partido Popular; PSOE: Partido Socialista Obrero Español; PD: Partito Democratico; Podemos: Podemos; SPD: Sozialdemokratische Partei; UMP: Union pour un mouvement populaire; UPyD: Unión Progreso y Democracia; UdC: Unione di Centro; UKIP: United Kingdom Independence Party.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Both authors participated in the design of the study, performed the statistical analysis, and drafted the manuscript. Both authors read and approved the final manuscript.

Authors' information

TY is a Research Fellow in Computational Social Science at the Oxford Internet Institute of University of Oxford. He is a physicist by training. JB is a Research Fellow at the Oxford Internet Institute of University of Oxford. He is a political scientist specialising in computational and 'big data' approaches to the social sciences.

Acknowledgements

This publication arises from research funded by the John Fell Oxford University Press (OUP) Research Fund; Grant Number: CZD06990.

Endnotes

[a] Other approaches such as socio-demographic and psychological approaches to voting offer less clear expectations about why people should seek information online.

[b] For example, the relevant page in the English version of Wikipedia for the 2009 EU Parliament elections is: http://en.wikipedia.org/wiki/European_Parliament_election,_2009.

[c] Which can be found at http://www.results-elections2014.eu/.

[d] Which can be found at http://www.nsd.uib.no/.

[e] Which can be found at http://www.lexisnexis.co.uk/.

[f] Diagnostic tests for these models (Tukey tests for non-additivity plus lack of fit tests for plots of Pearson residuals and Breusch-Pagan tests for heteroscedasticity) showed that improved versions of models 1.1 and 1.2 could be obtained by log transforming the dependent variable (vote share). These new models produced identical results to the original models in terms of the direction and statistical significance of effects, hence the original models have been reported to facilitate comparison and interpretation.

[g] Diagnostic tests for these models (Tukey tests for non-additivity plus lack of fit tests for plots of Pearson residuals and Breusch-Pagan tests for heteroscedasticity) showed that these models fitted the data well.

Received: 15 July 2015 Accepted: 5 June 2016 Published online: 18 June 2016

References

1. Lazer D, Pentland A, Adamic L, Aral S, Barabasi A-L, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M (2009) Life in the network: the coming age of computational social science. Science 323(5915):721-723
2. Margetts HZ, John P, Hale SA, Yasseri T (2015) Political turbulence: how social media shape collective action. Princeton University Press, Princeton
3. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with web search. Proc Natl Acad Sci USA 107(41):17486-17490
4. Curme C, Preis T, Stanley HE, Moat HS (2014) Quantifying the semantics of search behavior before stock market moves. Proc Natl Acad Sci USA 111(32):11600-11605
5. Kristoufek L (2013) Can Google trends search queries contribute to risk diversification? Sci Rep 3:2713
6. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012-1014
7. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, Del Valle SY (2014) Forecasting the 2013-2014 influenza season using Wikipedia. PLoS Comput Biol 11(5):e1004239
8. Mestyán M, Yasseri T, Kertész J (2013) Early prediction of movie box office success based on Wikipedia activity big data. PLoS ONE 8(8):e71226
9. Llorente A, Garcia-Herranz M, Cebrian M, Moro E (2014) Social media fingerprints of unemployment. PLoS ONE 10(5):e0128692
10. Yasseri T, Bright J (2014) Can electoral popularity be predicted using socially generated big data? Inf Technol 56(5):246-253
11. Metaxas PT, Mustafaraj E (2012) Social media and the elections. Science 338(6106):472-473
12. Tumasjan A, Sprenger T, Sandner P, Welpe I (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: ICWSM, pp 178-185
13. DiGrazia J, McKelvey K, Bollen J, Rojas F (2013) More tweets, more votes: social media as a quantitative indicator of political behavior. PLoS ONE 8(11):e79449
14. Gayo-Avello D, Metaxas P, Mustafaraj E (2011) Limits of electoral predictions using Twitter. In: ICWSM, pp 490-493
15. Mislove A, Lehmann S, Ahn Y, Onnela J, Rosenquist JN (2011) Understanding the demographics of Twitter users. In: Proceedings of the fifth international AAAI conference on weblogs and social media, pp 554-557
16. Jungherr A, Jurgens P, Schoen H (2011) Why the pirate party won the German election of 2009 or the trouble with predictions: a response to Tumasjan, A, Sprenger, TO, Sander, PG, & Welpe, IM 'Predicting elections with Twitter: what 140 characters reveal about political sentiment'. Soc Sci Comput Rev 30(2):229-234.
17. Lui C, Metaxas P, Mustafaraj E (2011) On the predictability of the US elections through search volume activity. In: Proc. IADIS int. conf., e-Society
18. Metaxas PT, Mustafaraj E, Gayo-Avello D (2011) How (not) to predict elections. In: Proceedings - 2011 IEEE international conference on privacy, security, risk and trust and IEEE international conference on social computing, PASSAT/SocialCom 2011, pp 165-171
19. Ofcom (2015) Adults' media use and attitudes report 2015
20. Rainie L, Tancer B (2007) Wikipedia users. Report
21. Weeks B, Southwell B (2010) The symbiosis of news coverage and aggregate online search behavior: Obama, rumors, and presidential politics. Mass Commun Soc 13(4):341-360
22. Curtice JK, Norris P (2008) Getting the message out: a two-step model of the role of the Internet in campaign communication flows during the 2005 British general election. J Inf Technol Polit 4(4):37-41
23. Schlozman KL, Verba S, Brady HE (2010) Weapon of the strong? Participatory inequality and the Internet. Perspective Polit 8(2):487-509
24. Fiske S, Taylor S (1991) Social cognition: from brains to culture. SAGE Publications, London
25. Utych SM, Kam CD (2014) Viability, information seeking, and vote choice. J Polit 76(1):152-166
26. Scheufle DA (2000) Twenty-five years of the spiral of silence: a conceptual review and empirical outlook. Int J Public Opin Res 12(1):3-28
27. Kaye BK, Johnson TJ (2002) Online and in the know: uses and gratifications of the web for political information. J Broadcast Electron Media 46(1):54-71
28. Dassonneville R, Dejaeghere Y (2014) Bridging the ideological space: a cross-national analysis of the distance of party switching. Eur J Polit Res 53(3):580-599

29. Dalton R (2013) The apartisan American. SAGE Publications, London
30. Chaffee S, Frank S (1996) How Americans get political information: print versus broadcast news. Ann Am Acad Polit Soc Sci 546(1):48-58
31. Chatterjee A, Mitrović M, Fortunato S (2013) Universality in voting behavior: an empirical analysis. Sci Rep 3:1049