Wikipedia and Political Candidates

Information Conditions during the 2022 United States Midterm Elections

Kaede Johnson, Shayan Khajehnouri, & Margaux Zwierski
EPFL Course DH-500

***Abstract*—We investigate 2022 US Midterm Candidates' Wikipedia pages to uncover de facto rules governing the relationship between a popular, "neutral" encyclopedia on the one hand and political candidates as well as Wikipedia editors on the other. Using logistic regression, OLS, and a host of descriptive statistics, we find proximity to political power is strongly related to trends in Wikipedia page ownership but insufficient in explaining the amount of edit activity a page receives. Additionally, trends in Wikipedia editors' edit comments reveal a general election-cycle edit timeline, while LDA applied to their edit content can surface an election's pressing issues - and perhaps a way to observe historical shifts in political favor.**

I. Introduction

Wikipedia is an online encyclopedia with content generated entirely by volunteers. In theory, anybody with an IP address is free to make edits on Wikipedia provided they adhere to the site's editing policies, which themselves stem from Wikipedia's "five pillars" - fundamental principles espousing neutrality, civility, public ownership, flexibility, and an encyclopedic style [1]. However, because neutrality can be a nebulous concept, flexibility provides coverage, and bad actors exist, editing barriers such as corrective bots [2] and edit access restrictions [3] have been erected.

Complications associated with neutrality requirements and restricted edit access are compounded in the case of events as highly publicized and politicized as elections. The 2022 American Midterm elections, with a large electorate observing a stable trend of extreme political division [4] and hundreds of Wikipedia pages across coinciding Senate, House, and Gubernatorial races, serve as a singular research opportunity for the relationship between Wikipedia and the political realm. While our original goal was to study ill-intentioned edit activity, Fahim et al. find such distortions on prominent politicians' Wikipedia pages are difficult to effect and even harder to see persist [5]. Thus, we take a more holistic approach, studying Wikipedia page presence and editor attention for politicians along with the amount of type of edits made by editors. In particular, we study the following four questions:

A. Which politicians have Wikipedia pages?

Agarwal et al. find massive peaks in visits to politician Wikipedia pages during election cycles, generally coinciding with engagement on other social media platforms [6]. In other words, readers visit Wikipedia at the same time they must decide which candidates to support. Lack of representation on Wikipedia, then, could compromise a candidate's ability to imprint his or her candidacy in readers' minds in a purportedly impartial setting. We are not aware of existing research on the question of Wikipedia page incidence among politicians, but there is similar research concerning the role of race and gender on Wikipedia page incidence (eg [7]).

B. Which politicians garner more Wikipedia page edits?

Following from the discussion in the previous subsection, more edits on a politician's Wikipedia page suggests his or her public-facing information flow during election season is more up-to-date than for other politicians.

On the other hand, because Borra et al. claim disputes about content in Wikipedia articles - for which we claim the number of edits in a normalized timeframe and candidate pool serves as a proxy - reflect larger societal debate [8], we can also frame this question as an investigation into our ability to predict which candidates are more likely to be part of the national discourse based on election-specific information.

C. Who is responsible for politicians' Wikipedia page edits?

Research has shown that a small proportion of editors are responsible for a high concentration of global edits on Wikipedia [9]. Observing this phenomenon in the political realm specifically would reveal an arbiter class of editors controlling information flow to potential voters. The balance of attention such editors afford different pages, meanwhile, may reveal discrepancies in arbiter class by political party or candidate.

D. What information receives edits on politicians' Wikipedia pages, and why?

The import of this analysis is evident: policy-related changes to public-facing information would (hopefully) matter more to potential voters than other changes. Priedhorsky et al. show that prolific editors also produce edits that are longer lasting (and therefore more influential) than less prolific editors [10]. It follows that researching differences in edits between editors would reveal what information is most influential.

II. Data

The basis for our data is the list of candidates who ran for Senator, Governor, or House Representative in the 2022 US Midterm elections. For all 1,318 candidates we transcribed the following attributes manually:

- Name
- Political Party
- Whether they were the incumbent
- Gender (based on name)[1]

[1] We consulted online photos where names were gender-neutral.

- The Cook Partisan Voting Index (CPVI) in the candidate's electoral region (this is one metric for race's competitiveness; see [11])
- Wikipedia page link, if available

Next, for all 638 available Wikipedia pages, we obtained the following data for the most recent 500 edits made:[2]

- Edit content (added or remove text)
- Date and time of edit
- Size of edit in bytes
- Editor comment regarding the reason for the edit
- Editor username; all usernames were pseudonymized before analysis.

III. Methods

Apart from descriptive analysis, we implement logistic regression, Ordinary Least Squares (OLS) regression, and Latent Dirichlet Allocation (LDA) to answer our research questions.

Both regressions incorporate the same set of explanatory variables: binary variables include status as incumbent, status as election winner, and status as male; categorical variables include intended office (Senate, House, or Governor) and political party (Democrat, Republican, or Other); and numerical variables incorporate the CPVI. To convert the CPVI into a numeric variable, we extract numerals from index values and multiply values from Democrat-preferring districts by -1.[3] A CPVI of EVEN is mapped to 0. $|\text{CPVI}|$, the absolute value of this numeric mapping, is our measure of a race's sheer competitiveness (lower = more competitive). We also interact the numeric mapping with binary variables indicating status as a Democrat or Republican, creating numerical measures for how preferred a candidate is based on their political party.

Our logistic regression regresses whether or not a 2022 US Midterm candidate **currently** has a Wikipedia page on the aforementioned explanatory variables (we also include a version where the dependent variable is whether or not a candidate has a Wikipedia page by November 8, 2022). Our OLS regression regresses log value of total 2022 edits made to a candidate's Wikipedia page on the aforementioned explanatory variables. In the latter regression, only those candidates with Wikipedia pages created before 2022 were retained. An overwhelming majority (over 90%) of candidates with Wikipedia pages meet this requirement (see Appendix Figure 1). Furthermore, solid variance in page creation dates even during 2022 (Appendix Figure 2) suggest a suitable level of variation to exploit for our logistic regression. Due to the skewed distribution in edits received by candidate, we excluded the top 10% most edited candidates from our OLS regression. In both regressions, we obtain regression coefficients using 90% of candidates' data, retaining 10% of candidates for model evaluation.

With LDA, we uncover latent topics present in the Wikipedia page edit content. Specifically, we access lines added or removed to Wikipedia pages between January 1, 2022 and now, splitting according to whether the edit happened before or after election day and on a Republican or Democratic candidate's page. We then preprocess this text by removing website links, Wikipedia formatting parameters, and punctuation as well as stemming all words, before implementing the LDA model from Python's gensim package [12] and extracting 10 topics for each intersection of the four intersections of political party and time period. By examining co-occurrences of patterns of words, we intend to identify key issues or discourse emerging during and after the 2022 Midterms.

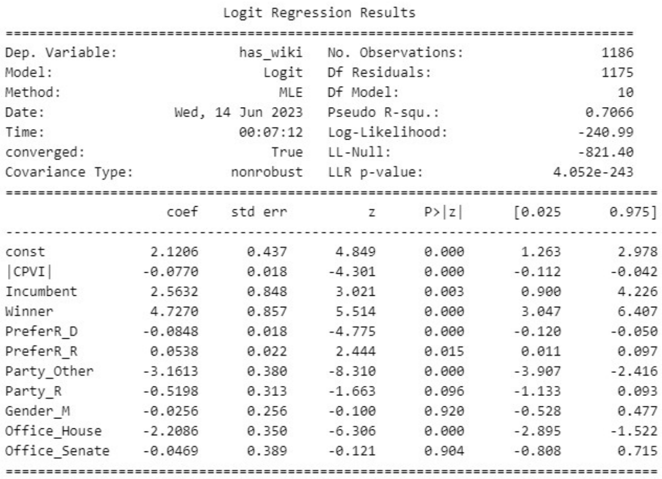
We conducted a series of descriptive analyses to support the quantitative methods mentioned above and dive further into the relationship between editors, candidates, and Wikipedia. These involved rank-order graphs, intra-editor edit attention distribution calculations, and aggregation of editor comments explaining their edits. Our was to "fill out" or understanding of edit activity and its implications for online discourse surrounding the election.

IV. Results & Discussion

A. Which politicians have Wikipedia pages?

The results of our logistic regression of Wikipedia page ownership on election and demographic factors are reported in Figure 1; the baseline candidate is a Democratic candidate for Governor. We find competitiveness ($|\text{CPVI}|$), status as incumbent, whether or not the candidate won, competitiveness interacted with political party (PreferR_D, PreferR_R), political party, and status as a House candidate to have a significant impact on a candidate's probability of having a Wikipedia page with p = .1. Applying our model on the 131 candidates reserved for the test set yielded 2 false positives and 8 false negatives - 92% accuracy, in other words (see Appendix Table I for the full confusion matrix).

To flesh out these regression results, we discuss trends for a select few variables. First, membership to the two major



```
                           Logit Regression Results
==============================================================================
Dep. Variable:               has_wiki   No. Observations:                 1186
Model:                          Logit   Df Residuals:                     1175
Method:                           MLE   Df Model:                           10
Date:                Wed, 14 Jun 2023   Pseudo R-squ.:                  0.7066
Time:                        00:07:12   Log-Likelihood:                -240.99
converged:                       True   LL-Null:                       -821.40
Covariance Type:            nonrobust   LLR p-value:                4.052e-243
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            2.1206      0.437      4.849      0.000       1.263       2.978
|CPVI|          -0.0770      0.018     -4.301      0.000      -0.112      -0.042
Incumbent        2.5632      0.848      3.021      0.003       0.900       4.226
Winner           4.7270      0.857      5.514      0.000       3.047       6.407
PreferR_D       -0.0848      0.018     -4.775      0.000      -0.120      -0.050
PreferR_R        0.0538      0.022      2.444      0.015       0.011       0.097
Party_Other     -3.1613      0.380     -8.310      0.000      -3.907      -2.416
Party_R         -0.5198      0.313     -1.663      0.096      -1.133       0.093
Gender_M        -0.0256      0.256     -0.100      0.920      -0.528       0.477
Office_House    -2.2086      0.350     -6.306      0.000      -2.895      -1.522
Office_Senate   -0.0469      0.389     -0.121      0.904      -0.808       0.715
================================================================================
```

Fig. 1. Logistic regression of Wikipedia page ownership on election and demographic factors. Baseline is a Democratic candidate for Governor. All variables that do not incorporate competitiveness are binary.

[2] Wikipedia's API limits edit histories to the 500 most recent edits. Only 16 candidates in our data had truncated 2022 edit histories due to this limit.

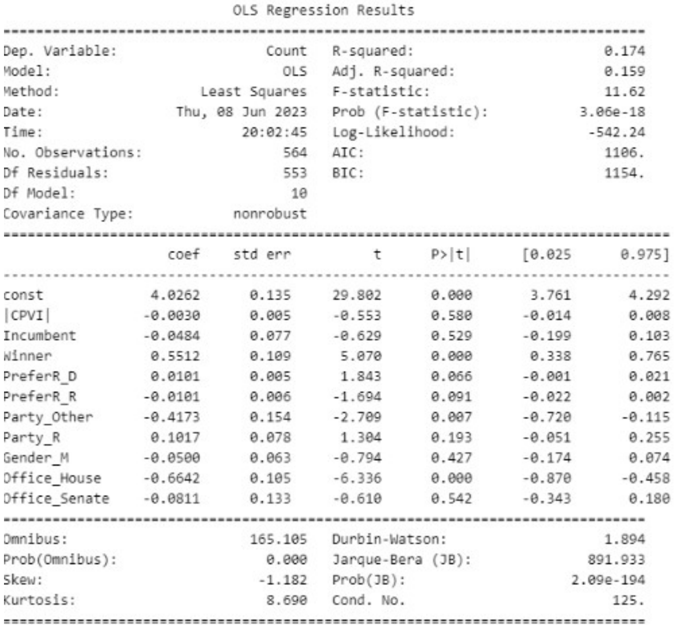
[3] Thus, a CPVI of D+4 corresponds to -4 while a CPVI of R+2 corresponds to 2.

US political parties is nearly essential for earning a Wikipedia page (see Appendix Figure 3). We interpret this result in two ways: (1) almost every winner in the US comes from one of these two parties, and it's only natural for an elected official to have a Wikipedia page, and (2) these two parties offer a notoriety boost even to non-winners that other parties simply cannot match. Indeed, just 5 of 129 Libertarian candidates - members of the country's third most popular political party - have Wikipedia pages, losing out to Independent candidates both in absolute and relative terms.

Also shown in the regression is sheer competitiveness' importance to Wikipedia page presence (see Appendix Figure 4 for a breakdown of Wikipedia presence among Senate candidates depending on race competitiveness; the picture for House candidates, available in Appendix Figure 5, is similar). We note that all Senate candidates in even or near-even races have Wikipedia pages, while the balance drifts toward 50-50 as elections become less competitive. Apart from editor opinions about the relationship between competitiveness and proximity to power, one possible reason for this phenomenon is that parties will run relatively unknown candidates in elections that are seemingly already lost, rendering editor attention even more unlikely.

Which politicians have Wikipedia pages? Those in power or whom editors believe will soon be in power. This explains why winning is the strongest signal for Wikipedia ownership (as measured by the magnitude of the coefficient) and why competitiveness is important even after controlling for political party, race type, and incumbency status. The interaction variables between competitiveness and party reaffirm this result, showing greater probability of winning increases the chance of having a Wikipedia page. One strange result is the decreased change of having a Wikipedia page associated with being a Republican; it is possible this indicates editors afford Democrats a lower threshold for "notability", though insignificance at the p = .05 and p = .01 levels mean this result should be viewed with a hint of caution. In fact, excluding whether or not the candidate wins their election from the regression and changing the dependent variable to whether or not the candidate has a Wikipedia page by election date (see Appendix Figure 6 and Table II for regression results and confusion matrix respectively) removes significant disparity between the two major political parties while maintaining the aforementioned results. In the case of this latter regression, it is truly editors' beliefs about the upcoming election which determine who earns a Wikipedia page, leading to serious gatekeeping in representation before the public and perhaps reinforcing the political status quo by lessening readers' exposure to unfamiliar candidates.

B. Which politicians garner more Wikipedia page edits?



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Count   R-squared:                       0.174
Model:                            OLS   Adj. R-squared:                  0.159
Method:                 Least Squares   F-statistic:                     11.62
Date:                Thu, 08 Jun 2023   Prob (F-statistic):           3.06e-18
Time:                        20:02:45   Log-Likelihood:                -542.24
No. Observations:                 564   AIC:                             1106.
Df Residuals:                     553   BIC:                             1154.
Df Model:                          10
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const             4.0262      0.135     29.802      0.000       3.761       4.292
|CPVI|           -0.0030      0.005     -0.553      0.580      -0.014       0.008
Incumbent        -0.0484      0.077     -0.629      0.529      -0.199       0.103
Winner            0.5512      0.109      5.070      0.000       0.338       0.765
PreferR_D         0.0101      0.005      1.843      0.066      -0.001       0.021
PreferR_R        -0.0101      0.006     -1.694      0.091      -0.022       0.002
Party_Other      -0.4173      0.154     -2.709      0.007      -0.720      -0.115
Party_R           0.1017      0.078      1.304      0.193      -0.051       0.255
Gender_M         -0.0500      0.063     -0.794      0.427      -0.174       0.074
Office_House     -0.6642      0.105     -6.336      0.000      -0.870      -0.458
Office_Senate    -0.0811      0.133     -0.610      0.542      -0.343       0.180
==============================================================================
Omnibus:                      165.105   Durbin-Watson:                   1.894
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              891.933
Skew:                          -1.182   Prob(JB):                    2.09e-194
Kurtosis:                       8.690   Cond. No.                         125.
==============================================================================
```

Fig. 2. Number of 2022 Wikipedia page edits regressed on election and demographic factors. The top 10% of candidates (as ranked by number of edits to Wikipedia page) were excluded, as were candidates without pages created before 2022.

The results of our OLS regression of log 2022 Wikipedia page edits on election and demographic factors are reported in Figure 2; the baseline candidate is a Democratic candidate for Governor. Both sheer competitiveness (|CPVI|) and incumbency status are no longer significant at the p = .1 level. Meanwhile, competitiveness interacted with political party (PreferR_D, PreferR_R) remains significant for Democrats but with a coefficient that has reversed its sign. Political party, win status, and status as a House candidate maintain the previous result. It is important to note that applying our model to the 56 candidates reserved for the test set yields a very weak positive correlation between actual and predicted 2022 edits counts (see Appendix Figure 7), suggesting the model is insufficient in explaining edit counts.

We again discuss trends for a few select variables to visualize these regression results. For example, edit counts between incumbents and non-incumbents were quite similar before election day in 2022 (see Appendix Figure 8), reaffirming that incumbent status alone is of little importance to number of edits received. Meanwhile, non-incumbents experience a hike in edits on election day that incumbents do not, reflecting a wave of non-incumbents "entering" the political class and showcasing winning's unequal effect on page attention.

Winning also has an unequal effect on the different candidate classes. While candidates for Senate, Governor, and House Representatives all receive similar edit rates in non-election weeks, candidates for House see effectively no boost on a normalized scale during election week despite an 8-to-10-fold spike for Senate and Governor candidates (see Figure 9). This may be due to the sheer number of house candidates; there are around ten times as many such candidates with Wikipedia pages then there are Senate or Governor candidates with Wikipedia pages).

Beyond the complexity of interaction in variables' impact on edit counts, perhaps the biggest takeaway from our OLS regression is the difficulty in predicting which candidates will receive edits based on the listed election factors alone. With just 4% of candidate Wikipedia pages receiving 20% of all 2022 edits, the edit count distribution is severely skewed (see Appendix Figure 10 for a rank-order graph of edits received in 2022), such that excluding the top 10% most-edited pages and log-transforming edit counts only pushes the regression's R-squared value up to .16. Potentially interesting interpretations - such as a sign reversal for PreferR_D suggesting Democratic candidates in unwelcoming districts inspire more attention or debate among editors - are sullied by the lack of strong explanatory power. With competitiveness seemingly irrelevant to edit counts, there remain no numeric explanatory variables in the regression, meaning there is little variation to exploit in trying to explain movements in a numeric variable.

With limited data augmentation capability (due to the need for a second round of ethics committee approval), the best way to further explain edit counts is by observing outliers. Those few candidates with more than 400 edits include Herschel Walker, Scott Jensen, Blake Masters, Mehmet Oz, and Doug Mastriano; common to all of them is the belief that the 2020 Presidential election was rife with improprieties damaging to Donald Trump's election prospects. Covid denialism and personal controversy related to abortion are also found without these candidates' Wikipedia pages.

Which politicians garner more Wikipedia page edits? Primarily those who are mouthpieces for the most contentious, truth-bending political topics, embroiled in controversy related to contentious political topics, or both. Regarding more "typical" candidates, while those who come in to power or who have the apparatus of a major political party behind them demonstrably receive more edits, incumbency and the sheer competitiveness of a candidate's race offer little signal. It would seem there are factors beyond electoral conditions affecting editors' attention, thwarting attempts to predict which candidates will receive up-to-date Wikipedia pages and/or become part of national discourse during the election cycle based on Wikipedia's data regarding electoral conditions alone.

C. Who is responsible for politicians' Wikipedia page edits?

Shifting attention to Wikipedia editors, we confirm that 2022 edits to Midterm candidates' Wikipedia pages were highly concentrated in a small subset of prolific editors. The skew in the distribution of edits made is stark: around 7,000 of the nearly 8,000 editors in our dataset made fewer than 10 edits in 2022, while the most prolific editors made hundreds, if not over 1,000 (see Appendix Figure 11 for a rank-order graph measuring edits made). That 0.003% of editors were responsible for 21% of all 2022 edits can be see in Figure 3. Here, we find that weekly edit activity among the most prolific editors peaks around the turn of July, during primary season. This stark jump is not seen among less prolific editors. Instead, the larger cohort sees peak edit activity around election day - the only time they are as active as the most prolific editors, however briefly (see Appendix Figure 12 for a non-cumulative look at edits-per-editor-cohort over time).

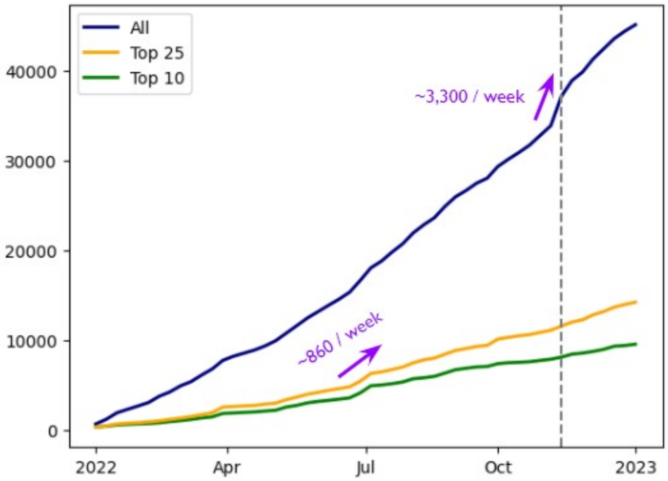


Fig. 3. Cumulative number of 2022 edits made based on activity cohort. Infrequent editors make the most edits on election day, while frequent editors make the most edits during primaries season.

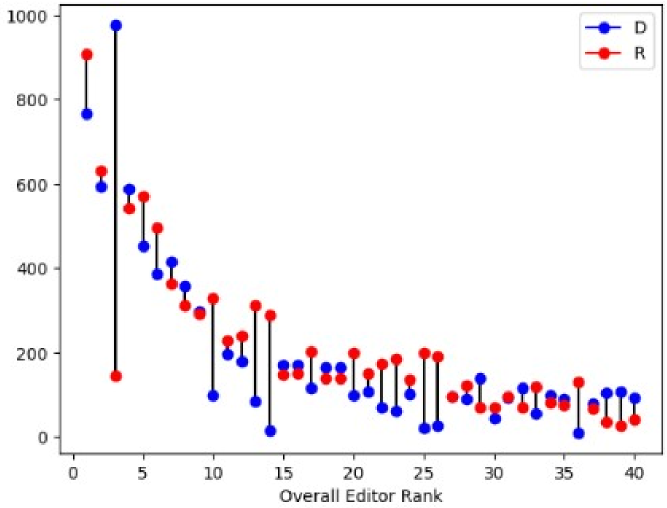


Fig. 4. Republican-to-Democratic-page distribution of edits made by top editors. Top editors generally balance their edits between the two major political parties.

Barring one major outlier and a handful of smaller gaps, intra-editor attention splits between Republican and Democrat Wikipedia pages appear relatively balanced among the most prolific editors in our dataset (Figure 4). The story is similar for slightly less prolific editors, though as one observes editors with tens (rather than hundreds) of total edits, instances of edits being entirely or almost entirely dedicated to one political party begin to proliferate (Appendix Figure 13). This effect pushes average attention toward Republican pages even if attention gaps are small in terms of number of edits; among the top 1000 editors by edit activity, just 265 made more edits on Democratic pages than Republican pages in 2022. This imbalance persists for other variables as well: 493 of the top 1000 editors dedicated more edits to female pages than male pages despite female candidates constituting less than 30% of the candidate set (see pertinent visuals in Appendix Figures 14 and 15).

Who is responsible for politicians' Wikipedia page edits? Practically, very few. These highly prolific "power" editors reach peak activity early in the election cycle - primary season and, to a lesser extent, early October - while less prolific editors seize on the set of "obvious" possible edits sprouting from election day to make safer changes that require less research. As the final election is the most well-known event in the election cycle, it is reasonable to assume less prolific editors either see opportunity to be "the first" to make important page updates or are driven by a desire to see their personal feelings about a race - positive or negative - reflected on a public-facing page about a candidate.

We also note that editors reveal an attention preference for Republican and Female pages through their edit activity. Rather than revealing tacit support, this could relate to increased incidences of editor "back-and-forths" on such pages. Such back-and-forth interactions likely proliferate on more popular pages given the rather linear trend between the number of edits on a given page and the number of distinct editors that have contributed to it (Appendix Figure 16). Even so, that it is a linear rather than strictly convex trend suggests back-and-forths are somewhat anchored by "power" users rather than one-off editors piling in to share their view. It would be interesting to see how observation variance proceeds beyond Wikipedia's 500-edit API limit. With current limitations, we can only note the revealed attention preferences, observe "power" user anchoring, and study with our next and final question the content of edits being made.

D. What information receives edits on politicians' Wikipedia pages, and why?

Our first approach to this question concerned editors own comments about their edits. Unfortunately, a plurality of such comments throughout 2022 were left blank. The next most popular edit comment, "External Links", describes little about the motivation for or implication of adding or removing specific links. Additionally, room for specificity means many thousands of comments were difficult to aggregate. Relaxed standards for editor comments therefore obfuscate the nature of many of the changes editors make (see Appendix Table III for an aggregated view of all 2022 edit comments).

We nonetheless graph the distribution of edit comments throughout 2022, aggregating into groups those comments attached to at least 100 revisions (Figure 5; note that nearly half of comments were not attached to more than 100 revisions and are therefore not included in the visualization). As expected, blank comments and comments mentioning "External Links" constitute a large share of edit comments, with election details the next most popular mention. Most interesting in this graph are trends for the "Stances - Social Policy" (teal ribbon), "Personal Life + Controversy" (pink ribbon), and "Blank" (bottom ribbon) categories. The Social Stances ribbon - in our view a proxy for all Stance edit activity due to the difficulties of aggregating non-social stance comments - peaks during June and July. Personal Life and Controversy edits, meanwhile, peak closer to election, in September and October.

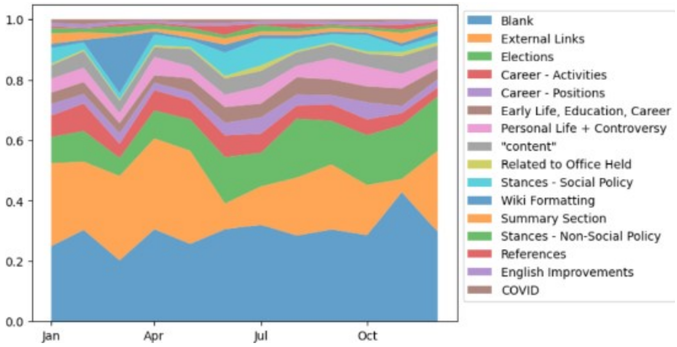


Fig. 5. Distribution of editor descriptions throughout 2022. Surges for political stance edits occur in the Summer, for personal life & controversy edits, in early Fall, and on election day for unexplained edits.

Finally, edits without editor comment peak during election month. We what uncover, then, is a life-cycle of election content according to the individuals responsible for edits.

To directly investigate editor impact without relying on editor statements, we report in Tables I and II the LDA-surfaced topics recovered from content added to or removed from Democratic and Republican pages respectively, considering edits made either before or after election date separately. We had hoped to see in these results clear indication of hot-button issues related to the upcoming election. One example is the "Healthcare" topic from Democrat-Before, which includes words related to abortion, along with the "Social Issues" topic in Republican Before, corresponding to abortion and gay marriage. We see in these two topics a desire to reaffirm political positions of each party. Discrepancies between the parties can be informative as well; "Defense" appears for Republicans but not Democrats, perhaps indicating increased resistance to ongoing defense spending in the Republican party.

Apart from these subjects, most recovered topics related to states important to each Party (New York and Maryland in the case of Democrats, Texas in the case of Republicans), current events (Biden, COVID19), or administrative happenings (Republican Caucus, House of Representatives, Political Administration). One time-based transition of note is the exchanging of Donald Trump in Republican Before for Ron DeSantis in Republican After - a potential herald of changes to the public's Republican front runner perception less than two years ahead of the 2024 Presidential election.

What information receives edits on politicians' Wikipedia pages, and why? We observe a general life-cycle of edits during the election, with policy-related edits occurring in Summer, edits related to personal or controversy occurring in early fall, and unexplained edits occurring during the election. Coupled with the discussion in the previous section, it is likely the most prolific editors - also the most influential [10] - who set information conditions for readers' exposure to policy stances. Election day, meanwhile, with an influx of one-off editors, a peak in unexplained edits, and a drastic dip in the average size of edits of edits made (see Appendix Figure 17) coupled with a surge in the size of total edits made (Appendix

Dem Before Election	Dem After
Election	?
New York	Maryland Gov.
Election	Michigan/Texas
?	Political Administration
Maryland Gov.	Covid19
Impeachment	Political Administration
Health	Impeachment
?	Election
Education	Biden
Georgia Election	Education

TABLE I
LDA PERFORMED ON CONTENT ADDED TO OR REMOVED FROM DEMOCRATIC CANDIDATES' PAGES BEFORE AND AFTER ELECTION DAY. PURPLE = NOT IN OPPOSITE COLUMN.

Rep Before Election	Rep After
?	Inauguration
Social Issued	DeSantis
Education	Senate Campagin
Texas/Supreme Court	?
Covid19	Rep. Caucus
Defense	Education
Election	Elections
Election	?
?	?
Donald Trump	House of Rep.

TABLE II
LDA PERFORMED ON CONTENT ADDED TO OR REMOVED FROM REPUBLICAN CANDIDATES' PAGES BEFORE AND AFTER ELECTION DAY. PURPLE = NOT IN OPPOSITE COLUMN.

Figure 18) is host to many small edits. As discussed above, and as confirmed by a similar dip in average edit size on the day the 118th Congress convened, these edits occur due to a tranche of new, important, and compact entering the public discourse. These differ from the sparser, larger edits made during the rest of the year.

Finally, LDA reveals it is hot-button issues, major states, current events, and administrative transition which received edit attention during the election. While differences between political parties and the periods before and after the election speak to differences in public discourse surrounding parties and candidates over time, these differences appear specific to a given election cycle. LDA on Wikipedia edit content, then, may be a valuable tool for making sense of public thought before, during, and after elections.

V. Conclusion

Performing a holistic analysis on Wikipedia pages for candidates in the 2022 US midterm elections, we find that in addition to a candidate winning an election, editors' beliefs about whether a candidate can win an election (if informed by competitiveness) will affects the candidate's chances of having a Wikipedia page. This explanatory power does not, however, extend to the number of edits a Wikipedia page receives, which appears more related to politicians' political stances or personal controversies - particularly if those political stances challenge the "reality" accepted by political opponents.

Regarding editors, we confirm the trend seen generally across Wikipedia of high edit concentration among a small cohort of editors. This despite investigating Wikipedia pages that are highly relevant to public life. Though generally balanced in relative terms among the top editors, attention does lean toward Republican and Female pages in the aggregate, perhaps due how these variables affect public discussion around a candidate. We also find that edit activity among prolific editors peaks during primary season, which is also when editor comments reveal a peak in edits related to candidates' political stances. Less prolific editors, meanwhile, make most of their edits on or near election day, likely seizing upon a plethora of newly available political happenings to make "safe" edits - although a coinciding spike in edits left unexplained by editors means these edits may reflect more than objective political updates.

Topic analysis was moderately successful at revealing issues important to the 2022 Midterm elections, though administrative details, large states' elections, and current events appear to crowd out policy-related topics. If this crowding out can be controlled, then applying LDA to edits would not only be helpful in identifying salient political topics, but it would also help trace movements in political thought. Better catering LDA to "see through" administrative details and populous states' elections is one way this capability can be further explored.

There are several ways to augment and scale up these research findings. First, researchers might incorporate more than the 2022 US Midterm elections into our methods, with fixed-effects regressions revealing variables ever-pertinent to Wikipedia ownership and activity as well as another dimension with which to observe stratification in topic analysis. Data outside of Wikipedia, such as candidate funding - excluded from this research due to ethics concerns - could improve our OLS regression by introducing more explanatory variance. Layering Google Trends data over the editor comment distribution timeline, meanwhile, would reveal the extent to which this timeline can affect readers' perceptions. Finally, developing a means of separating good-faith disagreement from obvious rule-breaking and trolling would allow our regressions and LDA to better target actual reflections of public thought. Ultimately, with research of greater scale, explanatory power, and targeting success, Wikipedia might be used a robust lens with which to understand the evolution of political thought among the public - and to understand how editors either reinforce or chip away at preconceived political notions.

References

[1] Wikipedia. Wikipedia:Five pillars — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Wikipedia%3AFive%20pillars&oldid=1159711137, 2023. [Online; accessed 13-June-2023].

[2] Lei Zheng, Christopher M Albano, Neev M Vora, Feng Mai, and Jeffrey V Nickerson. The roles bots play in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20, 2019.

[3] Benjamin Mako Hill and Aaron Shaw. Page protection: another missing dimension of wikipedia research. In *Proceedings of the 11th International Symposium on Open Collaboration*, pages 1–4, 2015.

[4] Gary C Jacobson. The 2022 elections: A test of democracy's resilience and the referendum theory of midterms. *Political Science Quarterly*, 138(1):1–22, 2023.

[5] Miles McCain Nick Rubin Maha Al Fahim, Sean Gallagher and Stanford Internet Observatory. Inauthentic editing: Changing wikipedia to win elections and influence people, 2001.

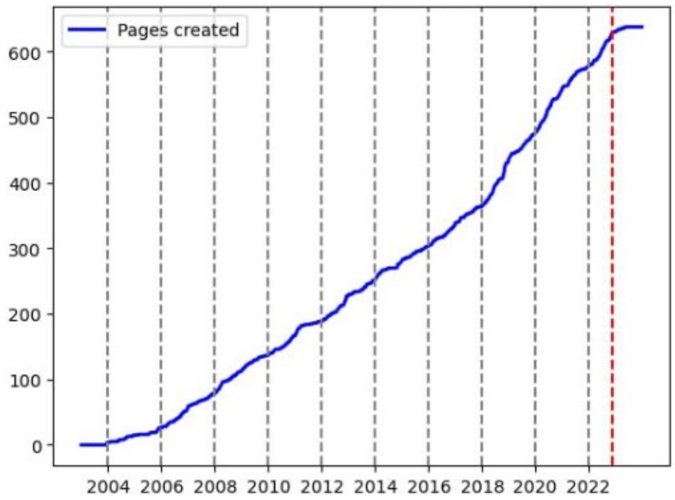


Fig. 1. Cumulative Wikipedia pages observed in our dataset since 2003. Two-thirds of candidates in our dataset had Wikipedia pages created before the 2018 Midterm elections.

[6] Pushkal Agarwal, Miriam Redi, Nishanth Sastry, Edward Wood, and Andrew Blick. Wikipedia and westminster: Quality and dynamics of wikipedia pages about uk politicians. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 161–166, 2020.

[7] Mackenzie Emily Lemieux, Rebecca Zhang, and Francesca Tripodi. "too soon" to count? how gender and race cloud notability considerations on wikipedia. *Big Data & Society*, 10(1):20539517231165490, 2023.

[8] Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. Societal controversies in wikipedia articles. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 193–196, 2015.

[9] Katherine Panciera, Aaron Halfaker, and Loren Terveen. Wikipedians are born, not made: a study of power editors on wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 51–60, 2009.

[10] Reid Priedhorsky, Jilin Chen, Shyong (Tony) K Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 259–268, 2007.

[11] JR. COOK, CHARLES E. Pvi. 2008.

[12] Radim Řehřek, Petr Sojka, et al. Gensim—statistical semantics in python. *Retrieved from genism. org*, 2011.

APPENDIX

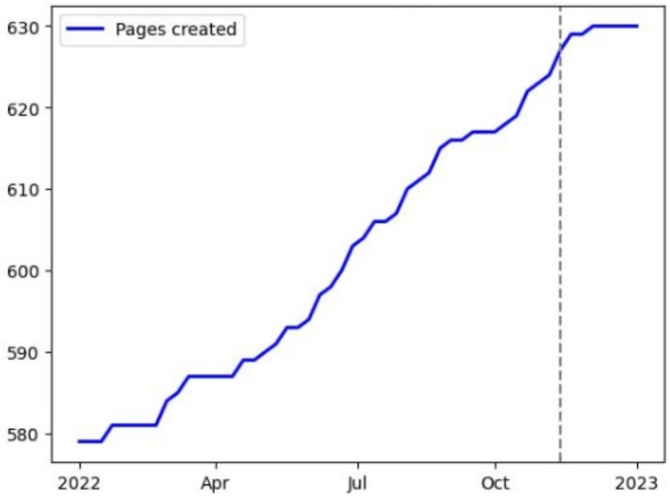


Fig. 2. Cumulative Wikipedia pages observed in our dataset throughout 2022. Candidates' Wikipedia pages were created steadily throughout the year.

		Prediction		
		No Wiki	**Wiki**	**Total**
Actual	**No Wiki**	64	2	66
	Wiki	8	57	65
	Total	72	59	

TABLE I
CONFUSION MATRIX FOR LOGISTIC REGRESSION OF WIKIPEDIA PAGE OWNERSHIP ON ELECTION AND DEMOGRAPHIC FACTORS. THE MODEL IS 93% ACCURATE ON THE 10% OF CANDIDATES RETAINED FOR MODEL EVALUATION.

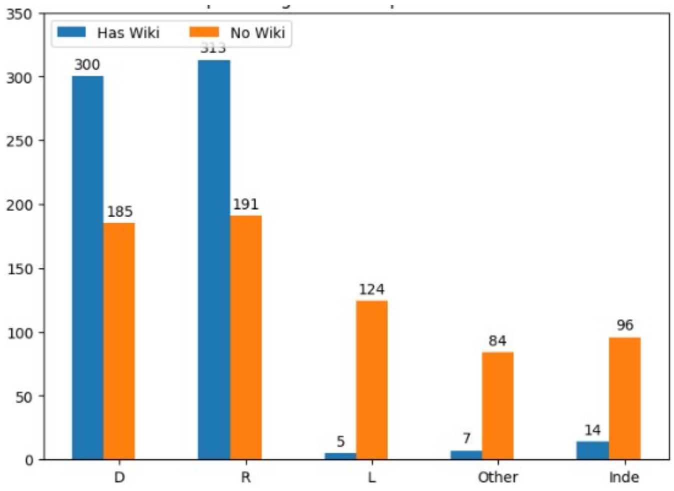


Fig. 3. Wikipedia ownership by candidate's political party. Political affiliation outside the Democratic or Republican parties is irrelevant to Wikipedia page ownership.

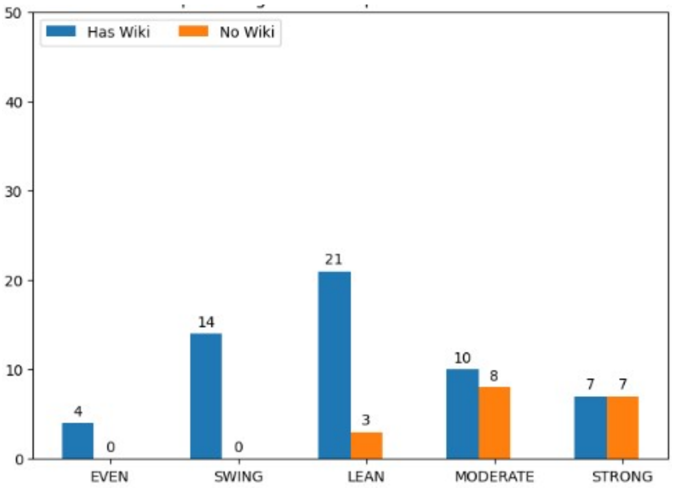


Fig. 4. Wikipedia ownership by competitiveness of Senate candidate's race. Candidate's in more competitive races are more likely to have Wikipedia pages.

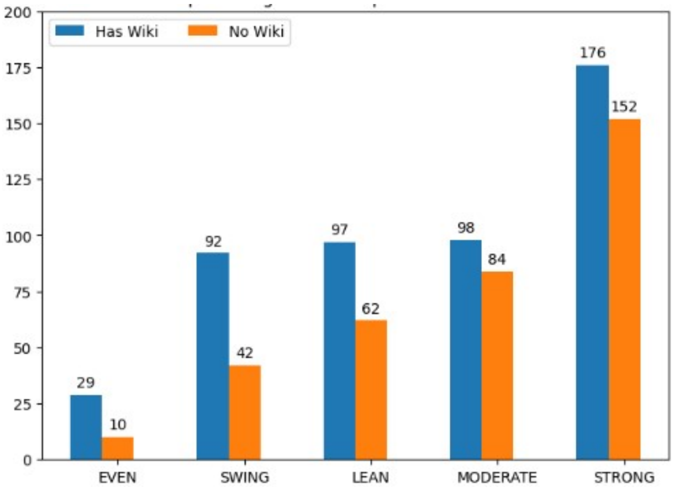
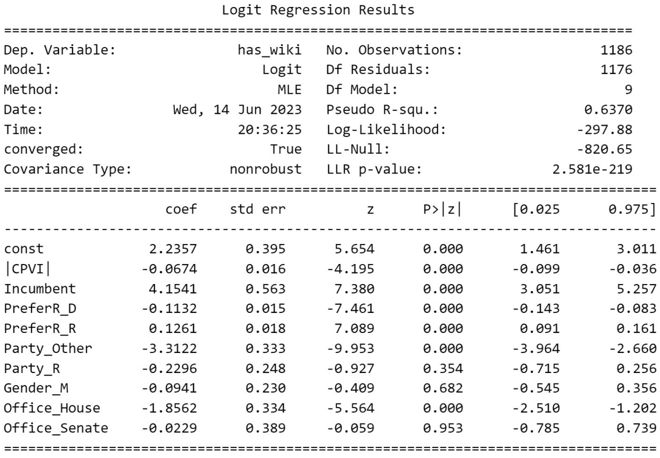


Fig. 5. Wikipedia ownership by competitiveness of House candidate's race. Candidate's in more competitive races are more likely to have Wikipedia pages.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:               has_wiki   No. Observations:                 1186
Model:                          Logit   Df Residuals:                     1176
Method:                           MLE   Df Model:                            9
Date:                Wed, 14 Jun 2023   Pseudo R-squ.:                  0.6370
Time:                        20:36:25   Log-Likelihood:                -297.88
converged:                       True   LL-Null:                       -820.65
Covariance Type:            nonrobust   LLR p-value:                 2.581e-219
=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const             2.2357      0.395      5.654      0.000       1.461       3.011
|CPVI|           -0.0674      0.016     -4.195      0.000      -0.099      -0.036
Incumbent         4.1541      0.563      7.380      0.000       3.051       5.257
PreferR_D        -0.1132      0.015     -7.461      0.000      -0.143      -0.083
PreferR_R         0.1261      0.018      7.089      0.000       0.091       0.161
Party_Other      -3.3122      0.333     -9.953      0.000      -3.964      -2.660
Party_R          -0.2296      0.248     -0.927      0.354      -0.715       0.256
Gender_M         -0.0941      0.230     -0.409      0.682      -0.545       0.356
Office_House     -1.8562      0.334     -5.564      0.000      -2.510      -1.202
Office_Senate    -0.0229      0.389     -0.059      0.953      -0.785       0.739
=================================================================================
```

Fig. 6. Logistic regression of Wikipedia page ownership on election and demographic factor (alternative version where dependent variable is whether the candidate had a Wikipedia page **by election date** and whether the candidate wins their election is removed from the explanatory variables). Baseline is a Democratic candidate for Governor. All variables that do not incorporate competitiveness are binary.

		Prediction		
		No Wiki	Wiki	Total
Actual	No Wiki	66	5	71
	Wiki	7	53	60
	Total	73	58	

TABLE II

CONFUSION MATRIX FOR LOGISTIC REGRESSION OF WIKIPEDIA PAGE OWNERSHIP ON ELECTION AND DEMOGRAPHIC FACTORS (ALTERNATIVE VERSION WHERE DEPENDENT VARIABLE IS WHETHER THE CANDIDATE HAD A WIKIPEDIA PAGE **BY ELECTION DATE** AND WHETHER THE CANDIDATE WINS THEIR ELECTION IS REMOVED FROM THE EXPLANATORY VARIABLES). THE MODEL IS 91% ACCURATE ON THE 10% OF CANDIDATES RETAINED FOR MODEL EVALUATION.

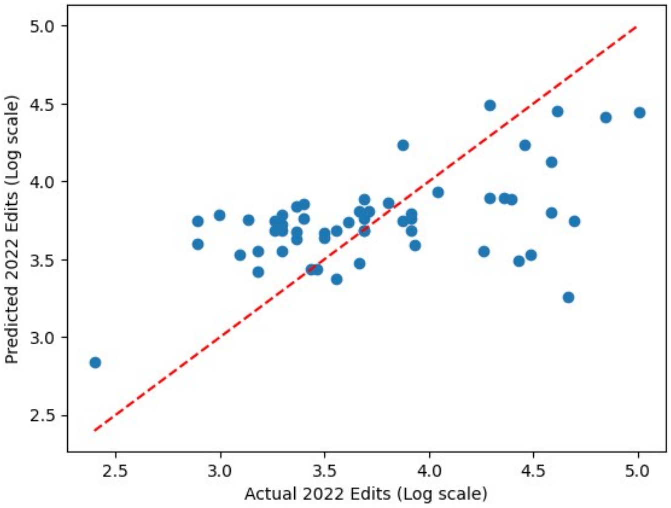


Fig. 7. Predicted number of 2022 Wikipedia page edits vs. Actual number of 2022 Wikipedia page edits for the 56 candidates reserved for OLS model testing. Small correlation between the two axes suggests the model's variables are insufficient in explaining edit rates.

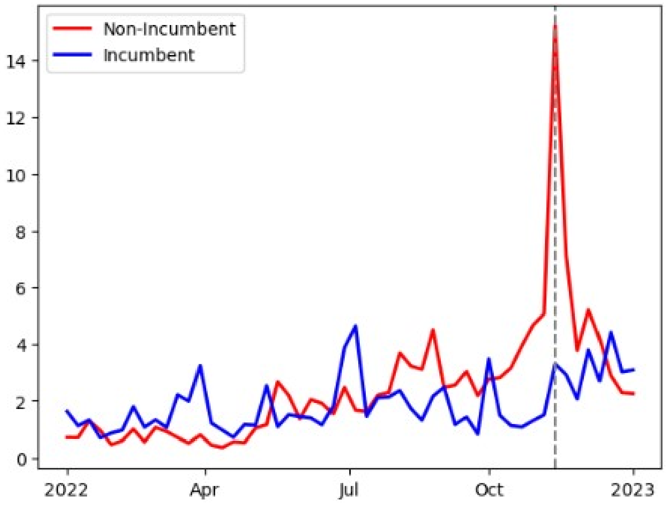


Fig. 8. Normalized edits received in 2022 by incumbency status. Winning matters more for non-incumbents than incumbents.

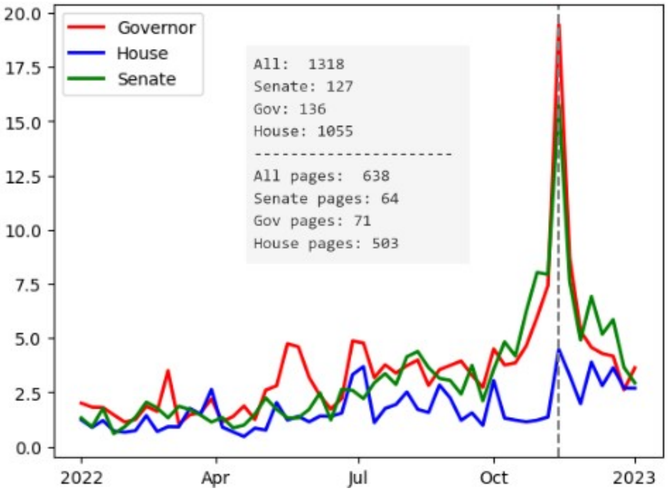


Fig. 9. Normalized edits received in 2022 by intended office. Senate and Governorship candidates receive more edits than House candidates - even on election day.

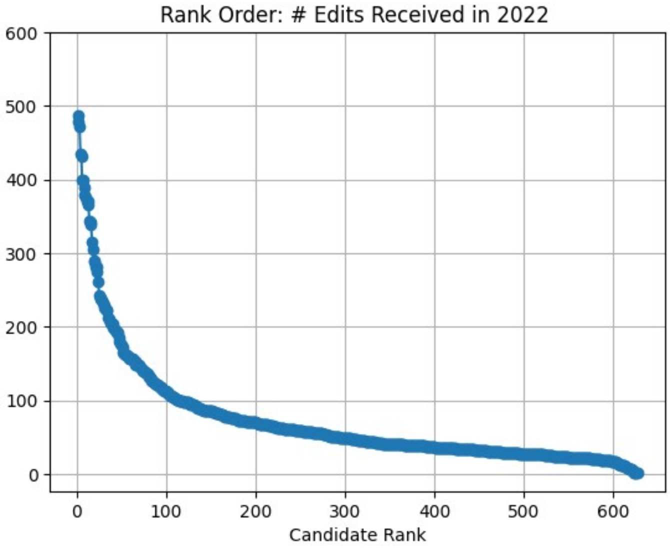


Fig. 10. Rank-order graph for number of Wikipedia page edits received in 2022. Edits are concentrated among a small cohort of candidates.

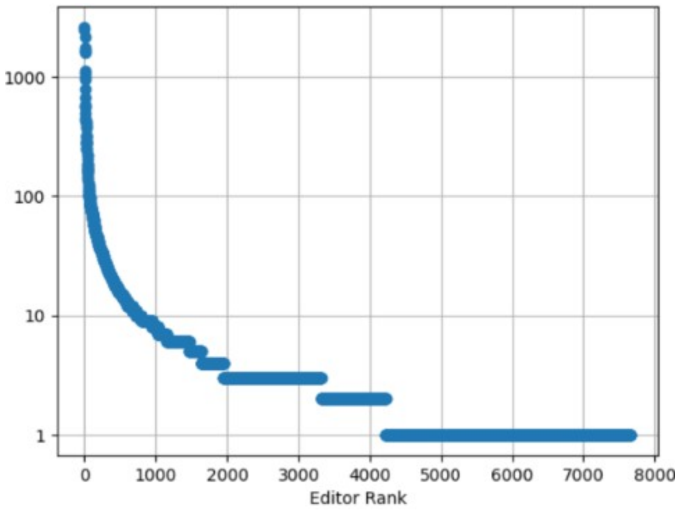


Fig. 11. Rank-order graph for number of Wikipedia page edits made in 2022. Edits are concentrated among a small cohort of editors

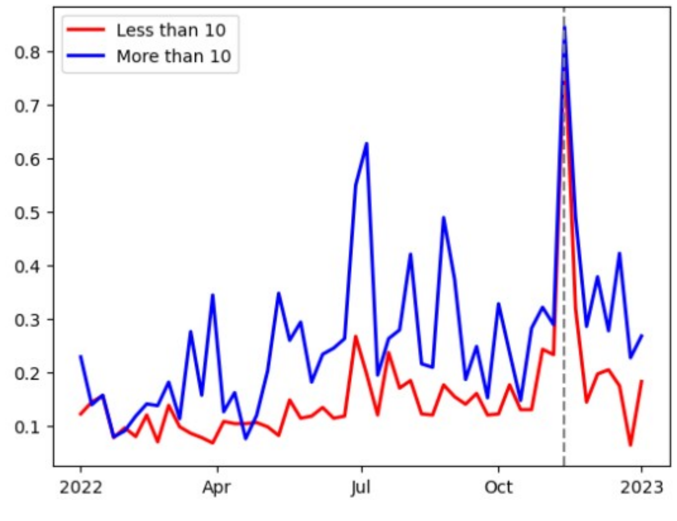


Fig. 12. Normalized number of 2022 edits made based on editor cohort. Infrequent editors make as many edits as frequent editors on election day.

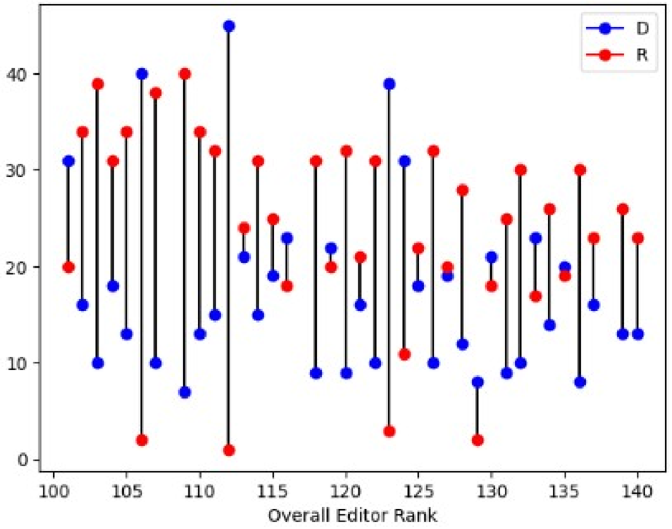


Fig. 13. Republican-to-Democratic-page distribution of edits made by frequent but not top editors. Such editors exhibit a wider relative gap between edits made to pages from the two major political parties.

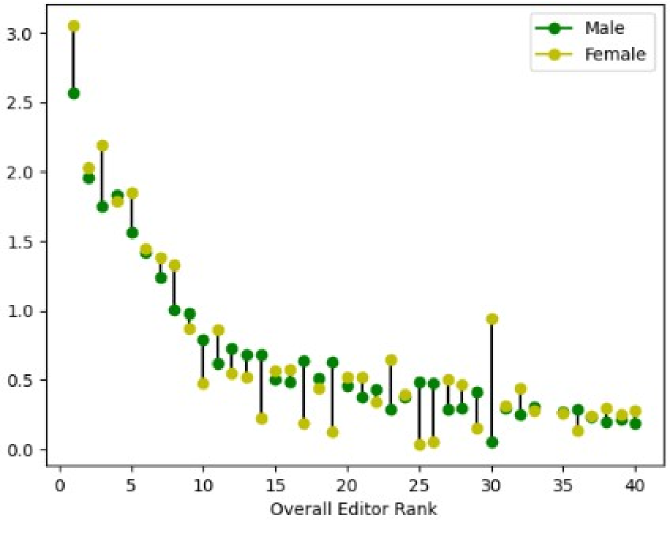


Fig. 14. Normalized male-to-female-page distribution of edits made by top editors. Top editors generally balance their edits between the two genders.

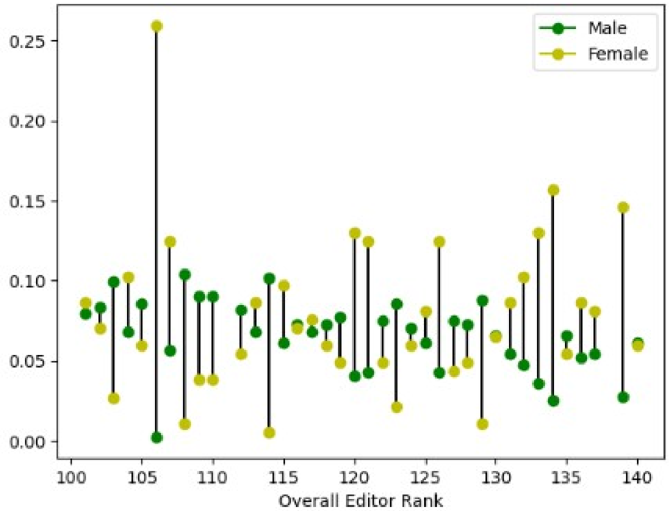


Fig. 15. Normalized male-to-female-page distribution of edits made by frequent but not top editors. Such editors generally balance their edits between the two genders.

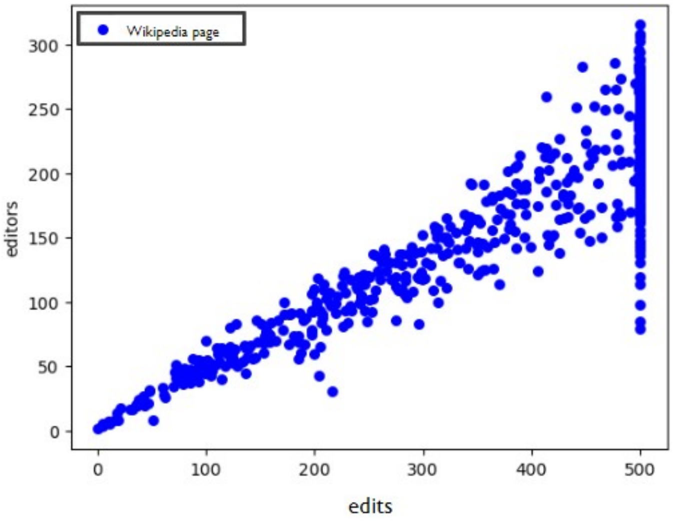


Fig. 16. Number of distinct editors on Wikipedia pages relative to number of 2022 page edits. The number of distinct editors generally increases linearly with edit size.

Revision Comment	%
	19%
External Link	10%
Elections	6%
Career - Activities	3%
Career - Positions	3%
Early life, Education, Career	3%
Personal Life + Controversy	2%
“Content”	2%
Related to Office Held	1%
Stances - Social Policy	1%
Wiki Formatting	1%
Summary Section	1%
Stances - Non-Social Policy	1%
References Section	1%
English Improvements	<1%
Covid	<1%
Unaggregated / Everything Else	**47%**

TABLE III
EDITOR DESCRIPTIONS OF THEIR EDIT CONTENT; COMMENTS ATTACHED TO AT LEAST 100 COMMENTS AGGREGATED INTO LARGER CATEGORIES. A PLURALITY OF EDITS RECEIVE NO EXPLANATION.

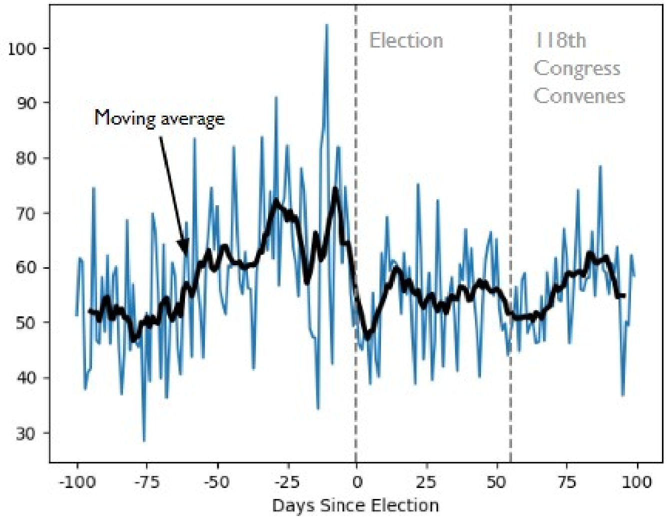


Fig. 17. Average size of edits over time. Dips on election day and the day Congress convenes reflect a surge in smaller edits.

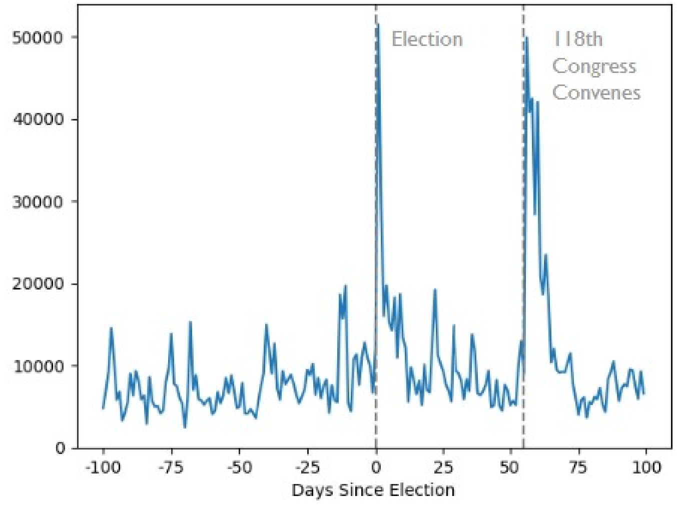


Fig. 18. Total size of edits over time. Surges on election day and the day Congress convenes reflect interest in reporting changes to the composition of Congress.