A Novel Methodology for Improving Election Poll Prediction Using Time-Aware Polling

Alexandru Topirceanu\textsuperscript{1*}, and Radu-Emil Precup\textsuperscript{2}
\textsuperscript{1}Department of Computer and Information Technology
\textsuperscript{2}Department of Automation and Applied Informatics
Politehnica University Timișoara, Romania
\{alext@cs.upt.ro*, radu.precup@aut.upt.ro\}

Abstract—Multiple poll forecasting solutions, based on statistics and economic indices, have been proposed over time, but, as we better understand diffusion phenomena, we know that temporal characteristics provide even more uncertainty. As such, current literature is not yet able to define truly reliable models for the evolution of political opinion, marketing preferences, or social unrest. Inspired by micro-scale opinion dynamics, we develop an original time-aware (TA) methodology which is able to improve the prediction of opinion distribution, by modeling opinion as a function which spikes up when opinion is expressed, and slowly dampens down otherwise. After a parametric analysis, we validate our TA method on survey data from the US presidential elections of 2012 and 2016. By comparing our time-aware method (TA) with classic survey averaging (SA), and cumulative vote counting (CC), we find our method is substantially closer to the real election outcomes. On average, we measure that SA is 6.3\% off, CC is 5.6\% off, while TA is only 1.5\% off from the final registered election outcomes; this difference translates into an \approx 75\% prediction improvement of our TA method. As our work falls in line with studies on the microscopic temporal dynamics of social networks, we find evidence of how macroscopic prediction can be improved using time-awareness.

Index Terms—election polls, time-aware, opinion prediction, social network mining.

I. INTRODUCTION

The modeling of diffusion processes can be inferred by designing interactions at microscopic level (i.e., between individual social agents), and forecasting network evolution at macroscopic level \cite{1, 2}. Namely, we often try to understand the macroscopic behavior by: (i) monitoring when social agents become indoctrinated by their neighborhood (i.e., they adopt information, get infected, buy merchandise \cite{3–5}), then, (ii) being able to predict how cascades of information flow, and eventually, how the diffusion process is percolated by individuals. Nevertheless, temporal aspects are shown to play an essential role in the diffusion of influence \cite{6, 7}. Many predictive assumptions are made only by observing when an agent gets indoctrinated, and not by considering the variable connectivity in the network \cite{8}, the dynamic trust that builds in ego-networks, or the sources of information \cite{9}.

Based on the available studies, we know for sure that social networks have a decisive role in the diffusion of information \cite{10}, and have proven to be very powerful in many situations involving macroscopic behavior \cite{11}. Examples include, but are not limited to, decisively influencing the Arab Spring in 2010 \cite{12}, and the U.S. presidential elections in 2008 and 2012 \cite{13}. The popularity of such online frameworks permits (most) people to spread information in a way that we can consider as new layer of social life \cite{7}.

Our study focuses on the predictability of election polls. This area of research was originally constructed by employing classic statistical models, applied on opinion polls prior to the election day \cite{14}. Ever since the late ’70s, it became a scientifically proposed fact that correct timing of the election date can be crucial for the outcome \cite{14}.

We build upon the premises to extrapolate macroscopic behavior of a society during the pre-election period, and make the following assumptions: (i) opinion poll results are made public before election day, and have a measurable effect on the whole social network (representing the voters); (ii) the results of each poll, as well as the timing of the polls, have a measurable effect on the opinion dynamics of the whole network; (iii) exposing voters to public opinion, through social media, acts similar to information cascades in the vicinity of a social agent; (iv) the temporal dynamics of an individual can be extrapolated to the temporal dynamics of the whole social network.

Many scientific models used in the prediction of diffusion phenomena revolve around the topological features of the underlying network, and are, thus, confined to one-dimensional property spaces \cite{7, 15}. In other words, these models intend to predict dynamical properties such as information cascade formation and population reach, but they ignore the temporal dimension. On the other hand, many temporal models define only microscopic behavior, without generalizing the impact on the whole network \cite{6}.

In this paper, we put several pieces of this puzzle together, as we consider the issue of predicting the temporal dynamics, and we propose a black-box prediction model which encapsulates the topological and behavioral properties of a network, all based on the properties of the microscopic level.
II. METHODS

When focusing on a prediction model where microscopic interactions are represented as information cascades triggered by so-called opinion sources (also spreader nodes, stubborn agents, vital nodes) \cite{5, 16, 17}. While these interactions are meaningful, accounting for all of them is still technologically, ethically, and legally impossible (e.g., analyzing all tweets posted in the USA during the pre-election period). As such, we use ubiquitous data under the form of opinion polls, which are centralized and open, on the Real Clear Politics \footnote{Available online at https://www.realclearpolitics.com} website. Specifically, we use the US Presidential Election polls from 2012 and 2016 for showcasing our prediction methodology.

We define the discrete temporal election axis \( t = [0, e] \) as being relative to the first opinion poll \( (p(0), t = 0) \), and the election day \( t = e \). The discrete observations we consider as opinion injection at any time \( 0 \leq t < e \) stem from all public opinion polls \( p(t) \) preceding the election day \( e \). All of the raw poll data have a timestamp under the form of start date–end date (e.g., ‘11/1 - 11/4’, the year is not provided, but we infer it from the election date). We consider only the end date (e.g., ‘11/1 - 11/4’), then \( e \) remains unchanged, so that, as time increases \( t \) increases, \( \alpha_t \) will decrease; if there exists a poll \( p_i(t) > 0 \) at the current moment, then \( \alpha_t \) is increased by an amplitude proportional to the number of votes (or normalized number of votes). The evolution of \( \alpha_t \) is given by the following equation:

\[
\alpha_t = \begin{cases} 
\alpha_t(t-1)t^{-\beta_t} + p_i^*(t), & \text{if } \exists p_i(t) > 0 \\
\alpha(t-1), & \text{if } p_i(t) = 0
\end{cases}
\]

Here, \( p_i^*(t) \) represents the normalized number of votes expressed in support for opinion \( i \) at time \( t \). In order to handle variation in the amplitude of polls, we normalize each value in the range \([0, 100]\%\) of the total expressed opinion at time \( t \). In addition to our proposed time-aware method, we use cumulative counting (CC) and survey averaging (SA) to serve as basic statistical methods for comparison.

The CC method is applied by summing up all votes expressed by the polls \( p_i \) for each opinion \( i \) over the total polling period \([0, e]\). Here, we also define a cumulative weight \( cw_i(t) \) for each opinion \( i \) which is updated as:

\[
cw_i(t) = \begin{cases} 
cw_i(t-1) + p_i(t), & \text{if } \exists p_i(t) > 0 \\
cw_i(t-1), & \text{if } p_i(t) = 0
\end{cases}
\]

The SA method is applied by averaging the current normalized poll results with the previously computed average, over each independent opinion. Here, we can express the opinion poll \( s\Omega_i \) directly by using the normalized (*) number of votes for each poll \( p_i^* \):

\[
s\Omega_i(t | 0 \leq k \leq t) = \frac{\sum_k p_i^*(k)}{|\{p_i^*(k) | 0 \leq k \leq t\} |}
\]

In Equation 3 we obtain the current poll at time \( t \) by summing up all normalized votes for the period \([0, t]\) (hence \( k \leq t \)) and divide by the number (cardinal) of polls in that same period.

III. RESULTS

Given the political system in the USA, both datasets used for validation contain opinion expressed for the Democratic Party candidate (2012: Barack Obama, 2016: Hillary Clinton), Republican Party candidate (2012: Mitt Romney, 2016: Donald Trump), and for other candidates grouped together in the “Others” category. We use as ground truth for validation the actual poll results from each respective election.

In Figure 1 we exemplify how CC and TA work, by plotting the respective weights \( cw_i(t) \) and \( w_i(t) \) for the final period before election (last 150-200 days). Figure 1 shows a distinctive macroscopic dynamics of the weights \( w_i(t) \) for each opinion that is shaped by the timing of each published poll. Specifically, each spike in Figure 1 corresponds to a poll being made public at time \( t \); the period in between two polls, corresponds to a damping of the weights in the same plot. Our assumption is that this time-aware rhythmic rise and fall of the weights better corresponds to the real-time inertia of each opinion.

Next, in Figure 2 we infer the evolution of actual poll results, in time, based on the measured weights from Figure 1. The snapshot in Figure 2 shows how the CC method
prediction is quite static, as it clearly suggests an advantage for the democratic (blue) candidate in 2012 (i.e. Clinton). The variation in predicted polls remains within large limits of > 5%. Nevertheless, Figure 2b shows a much more dynamic prediction system where the advantage of the democratic candidate (blue) ranges between [-0.94–4.90]% over the last 100 days before election.

In Table I we provide the real and predicted poll results based on the two available validation datasets, using the cumulative counting (CC), survey averaging (SA), and the two implemented TA methods: power-law time-aware (PA), and exponential time-aware (EA) polling.

The performances of all poll prediction estimation methods are displayed in Figure 3. We highlight in Figures 3a,b the results for the Democratic candidate (blue, Obama), respectively the Republican candidate (orange, Romney) corresponding to the 2012 elections. Similarly, in Figures 3c,d, we highlight the results for the 2016 Democratic (blue, Clinton) and Republican (orange, Trump) candidates. The first column in each panel represents the ground truth value, i.e., the real election results. The superior prediction power of PA and EA becomes visible, both visually as well as numerically. For the 2012 democratic candidate (Obama) we measure offsets of 2.13–2.18% for the TA methods, while CC and SA are offset by > 4%; for the Republican candidate (Romney) we measure offsets of only 0.35–0.42% for the TA methods, and CC and SA are offset by ≈ 3%. For the 2016 Democratic candidate (Clinton) we measure very small offsets of 0.14–0.16% for the TA methods, while CC and SA are offset by 0.51–1.39%; for the Republican candidate (Trump) the offsets are negligible (0.02–0.04%).

**Table I: Predicted and Real Poll Results for the 2012 and 2016 US Presidential Elections, Given in Percentages (%).**

<table>
<thead>
<tr>
<th></th>
<th>2012 Elections</th>
<th>Democratic</th>
<th>Republican</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real results</td>
<td>51.10</td>
<td>47.20</td>
<td>1.70</td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>46.77</td>
<td>44.23</td>
<td>9.00</td>
<td></td>
</tr>
<tr>
<td>SA</td>
<td>47.06</td>
<td>44.36</td>
<td>8.58</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>48.92</td>
<td>47.55</td>
<td>3.53</td>
<td></td>
</tr>
<tr>
<td>EA</td>
<td>48.97</td>
<td>47.62</td>
<td>3.41</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2016 Elections</th>
<th>Democratic</th>
<th>Republican</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real results</td>
<td>48.20</td>
<td>46.10</td>
<td>5.70</td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>47.69</td>
<td>42.70</td>
<td>9.61</td>
<td></td>
</tr>
<tr>
<td>SA</td>
<td>46.81</td>
<td>41.77</td>
<td>11.42</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>48.36</td>
<td>44.67</td>
<td>6.97</td>
<td></td>
</tr>
<tr>
<td>EA</td>
<td>48.34</td>
<td>44.65</td>
<td>7.01</td>
<td></td>
</tr>
</tbody>
</table>
candidate (Trump) we measure offsets of 1.43-1.45% for the TA methods, and CC and SA are offset by 3.40-4.33%.

IV. CONCLUSION

Inspired by temporal models for infectious contagion [20], [21], and dynamic diffusion models [5], we propose the idea of extrapolating the macroscopic behavior of a social network using time-aware microscopic models. As such, we implement a time-aware poll prediction method using the power-law (PA) and exponential (EA) models which are sensitive to the timing of opinion injection in a closed system of voters.

Despite the simple assumptions, our results pinpoint to the fact that time-awareness is more significant in poll prediction performance than previously considered. For the 2012 elections, we are able to approximate the final results within a 2% margin, while classic methods like SA and CC produce offsets of about 7%. Similarly, for the 2016 elections, our method manages to come within 1.5% of the real election results, while SA and CC stay outside the 4% margin. In terms of quantifying the overall performance boost of our method, compared to the benchmark methods, TA proves to be ≈ 75% more accurate for the 2012 elections, respectively ≈ 74% for the 2016 elections.

Current solutions for prediction employed by respectable institutions in the US, like the Huffington Post, Real Clear Politics, or Five Thirty Eight, employ poll counting and combining polls with economic indices. Nevertheless, we have not seen any time-aware method that is similar to the one proposed in this paper.

ACKNOWLEDGMENT

Authors A.T. and R.-E.P. are supported by the Romanian National Authority for Scientific Research and Innovation (UEFISCDI), project number PN-III-P1-1.1-PD-2016-0193.

REFERENCES


