Depolarization of opinions on social networks through random nudges

Ritam Pal,* Aanjaneya Kumar,† and M. S. Santhanam‡
Department of Physics, Indian Institute of Science Education and Research, Dr. Homi Bhabha Road, Pune 411008, India

(Received 23 February 2023; accepted 17 August 2023; published 15 September 2023)

Polarization of opinions has been empirically noted in many online social network platforms. Traditional models of opinion dynamics, based on statistical physics principles, do not account for the emergence of polarization and echo chambers in online network platforms. A recently introduced opinion dynamics model that incorporates the homophily factor—the tendency of agents to connect with those holding similar opinions as their own—captures polarization and echo chamber effects. In this work, we provide a nonintrusive framework for mildly nudging agents in an online community to form random connections. This is shown to lead to significant depolarization of opinions and decrease the echo chamber effects. Though a mild nudge effectively avoids polarization, overdoing this leads to another undesirable effect, namely, radicalization. Further, we obtain the optimal nudge probability to avoid the extremes of polarization and radicalization outcomes.

DOI: 10.1103/PhysRevE.108.034307

I. INTRODUCTION

The information revolution has lowered the entry barrier for nearly everyone to participate and contribute to shaping opinions and policies on various issues. This has been largely aided by the easy availability of social media infrastructure through mobile devices. Increasingly, the collective opinions expressed through various social media platforms are thought to be one barometer of the public mood on any contentious issue of the day [1]. This provides an interesting testing ground for the dynamics and statistical physics of interacting multiagent systems since the online nature of interactions provides fine-grained data for quantitative analysis and comparison with model results.

The study of opinion formation and its dynamics has attracted researchers for decades. The analysis of opinion dynamics from the statistical physics perspective can be traced back to the work of DeGroot [2], which provides a framework for reaching a consensus. Other discrete models, including the voter [3,4] model, Sznajd model [5,6], and their variants which have a strong basis in a framework of interacting spins, suggest that large participatory interactions among agents might also lead to the emergence of consensus. However, empirical results have shown that the distribution of opinions tends to show a bimodal distribution pattern corresponding to polarization, especially on controversial issues of the day [7–9]. Culture dissemination model [10], one of the first higher-dimensional modeling approaches to opinion dynamics, which also incorporates the human tendency to interact with similar persons, shows that despite there being local convergence, global polarization can be reached. Other discrete models [11–14] explain the effects of consensus, attitude changes in groups, and the spreading of minority opinions. In the presence of stubborn agents, these models can also capture the effect of polarization [15–17]. Different variants of the bounded confidence model [18,19] can also capture many empirically found trends in the distribution of opinions. These models can reproduce consensus, bimodal, or multi-modal opinion distributions depending on the confidence interval.

Another empirical feature that could not be accounted for by early models (at least by their original versions) was the phenomenon of echo chambers [20]. This refers to a scenario in which one agent's opinion is similar to the agents in their "social neighborhood," and one tends to reinforce the other. Lack of sufficient engagement with opposing opinions leads to positive reinforcement of one's own opinion within a close-knit social network. Empirical evidence for this effect has been reported from several social media platforms [21–24]. Few recent opinion dynamics models [25–28] have qualitatively captured the features of echo chambers, which have been shown to arise from personalized interactions among peers in an online setting, which might be accelerated through the platform's recommendation engine.

The model introduced by Baumann *et al.* accounted for several observed features from empirical data along with echo chambers in social media. The features that (a) most active users tend to be strongly opinionated and (b) locally connected agents have a convergence of opinions can be linked to the mechanism of reinforcement of opinion among agents and the tendency of agents to interact more with those with similar opinions (homophily [29,30]). Even if the model starts from an initial distribution of opinions without clear preferences, highly homophilic interactions induce the formation of echo chambers and polarized states.

Though having diverse opinions might be a desired outcome, extreme polarization leads to network segregation [31], which often bottlenecks the information flow in social networks. Also, echo chambers, often linked to polarization, are known to be responsible for sustaining misinformation for

*ritam.pal@students.iiserpune.ac.in
†kumar.aanjaneya@students.iiserpune.ac.in
‡santh@iiserpune.ac.in

a longer time on social networks [32,33]. These problems call for intervention mechanisms, which should be safe and noninvasive.

It might appear that in the case of controversial topics, the interaction and the debate will always lead to polarized states of opinion. But the underlying mechanism for polarization, the reinforcement of opinions through interaction between like-minded people, leaves us wondering if any intervention will help to reconcile disparate opinions.

In this work, we show that if agents are nudged slightly, then the cycle of reinforcement of opinions can be broken, and depolarization can be achieved. In social networks, the nudges are effected by exposing the agents to diverse opinions. We also show that overdoing this leads to radicalization [34,35], a state where all the agents have the same stance on an issue. We formulate an optimization problem that avoids polarization and radicalization and computes the right amount of nudge probability required to achieve this optimal scenario.

In the next section, we discuss the basic model and motivate the random nudges in the subsequent section. In Sec. IV, we demonstrate our results and discuss their implications. We formulate an optimization problem in Sec. V, which emerges from a tradeoff between depolarization due to the proposed random nudges and the tendency to move toward a radicalized state. We conclude with a discussion of future directions.

II. BASIC MODEL AND METHODS

To analyze polarization and to introduce possible intervention methods for reducing polarization, we adapt a recently introduced model for opinion dynamics [25]. This model qualitatively captures a few aspects of opinion dynamics when agents' opinions evolve due to interactions in social media platforms. The model can reproduce the empirical features such as polarization and echo chambers and the fact that more active people on social media tend to have extreme opinions.

The model has N interacting agents, and it is assumed there are only two possible sides to an issue. This is typical of many, but not all, the issues—for example, to allow abortion or not. Opinion on a given issue is denoted by x_i, which can take any real value in the range $(-\infty, \infty)$. The sign of the x_i corresponds to the stance of the agent in the corresponding issue, and $|x_i|$ denotes the conviction of the agent in their respective stance. This implies that the larger the value of $|x_i|$, the more extreme the agent's opinion is. The model used to capture the evolution of opinion is activity driven [36–39], i.e., at each time step, only active agents can influence other agents. Based on empirical data [36,38], the distribution of agent's activity is chosen to be

$$F(a) = \frac{1-\gamma}{1-\varepsilon^{1-\gamma}} a^{-\gamma}, \tag{1}$$

where a is the activity, ε is the minimum activity (chosen in this work to be 10^{-2}), and γ controls how steep the function $F(a)$. It is chosen to be $\gamma = 2.1$. Agents' opinions evolve based on their interactions with other agents, and this information is encoded in the time-dependent adjacency matrix $A_{i,j}(t)$. Further, opinion evolution also depends on the strength of social interaction $K > 0$ and the controversialness of the issue $\alpha > 0$. The opinion dynamics is given by the following N coupled differential equation [25]:

$$\dot{x}_i = -x_i + K\left(\sum_{j=1}^{N} A_{ij}(t)\tanh(\alpha x_j)\right). \tag{2}$$

In this, $A_{i,j}(t)$ is the temporal adjacency matrix of interaction at time t. If at time t agent j influences agent i, then $A_{i,j}(t) = 1$, and $A_{i,j}(t) = 0$ otherwise. If agent i is active at time t, they will interact with m other agents, weighted by the probability $P_{i,j}$. Further, the probabilistic reciprocity factor $r \in [0, 1]$ determines the chance that an interaction is mutually influential, i.e., $A_{ij}(t) = A_{ji}(t) = 1$. The interaction probability is defined to be a function of the magnitude between two agents' opinions:

$$P_{ij} = \frac{|x_i - x_j|^{-\beta}}{\sum_k |x_i - x_k|^{-\beta}}, \tag{3}$$

where β is the homophily factor which quantifies the tendency for agents with similar opinions to interact with each other; $\beta = 0$ refers to the absence of interaction preference; and $\beta > 0$ implies that the agents with similar opinions are more likely to interact with one another. Evidently, Eq. (3) is modeled as a power-law decay of connection probabilities with only a small chance for agents with opposite opinions to interact. Since most of the interactions tend to occur between agents with similar opinions, this can lead to the formation of echo chambers.

The interaction dynamics in the model is enforced by the activity-driven temporal network that is fully encoded by the parameters $(\varepsilon, \gamma, m, \beta, r)$, together with the parameters that characterize the issue, (K, α). Asymptotically, this model features three distinct states in the distribution of opinions. If the social interaction K is sufficiently small, then the opinion of every agent decays to zero, and this state is known as the neutral consensus state. However, if social interaction K is large, but the homophily factor β is small, then, due to statistical fluctuations, all the opinions either become positive or negative. This state, where each agent has the same stance (the sign of x_i for all i is the same) with possibly different convictions, is called radicalization. It is important to note that radicalization is an absorbing state of this model. This is because when all agents have opinions with the same sign, the dynamics does not allow for a sign change of any agent's opinion. The most interesting case emerges when social interaction K and homophily factor β are large enough. In this case, a metastable polarized state emerges, which is characterized by a bimodal opinion distribution.

III. RANDOM NUDGES AND POLARIZATION

Echo chambers are increasingly becoming more apparent in online social media platforms. A generic tendency to interact with people who hold similar opinions as ours can lead to echo chambers, and this effect is, in turn, amplified by the recommendation engines on social media platforms. These algorithmically driven engines recommend similar connections or content in order to keep the users of those platforms engaged.

These two features are modeled by the interaction probability, controlled by the homophily factor β. Large values

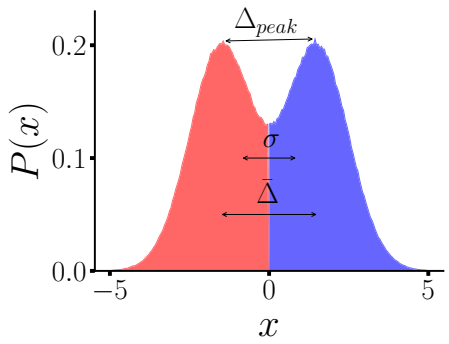

FIG. 1. A schematic to illustrate three measures of polarization. $\bar{\Delta}$ is the distance between mean positive and negative opinions. Δ_{peak} denotes the distance between two peaks in the opinion distribution, and $\boldsymbol{\sigma}$ denotes the standard deviation of the opinion distribution.

of β represent how closed the echo chambers are. To disrupt the formation of echo chambers, even while keeping the platforms as engaging as possible and without violating the users' privacy, we adopt the following intervention in the opinion dynamics model: With probability $p < 1$, the active agents will interact uniformly with any other agents, and with probability $(1 - p)$, the active agents will interact with others according to the homophily probability given in Eq. (3). We call p the random nudge probability. As p does not depend on the opinions of the agents, the intervention is noninvasive (the recommendation engine need not interpret the opinion of the agents). For small enough values of p, it is hoped that the platform is still engaging while maintaining enough diversity to ensure there is no echo chamber. With this intervention, we propose a modified interaction probability as

$$\widetilde{P}_{ij} = p \times \frac{1}{N-1} + (1-p) \times P_{ij}. \tag{4}$$

This is used in the rest of the results shown in this paper.

Quantifying Polarization. Before we delve into the details of the results, we discuss the three quantities employed to measure the degree of polarization based on the opinion distribution $P(x)$. They are defined as (a) Polarization measured through $\bar{\Delta}$, defined as the distance between the average of positive opinions and the average of negative opinions. (b) When opinion distribution exhibits a bimodal character, the distance between the two peaks, denoted by Δ_{peak}, can also be used as a measure of polarization [41]. (c) A gross measure of polarization could also be the standard deviation σ of the entire opinion distribution [27]. Figure 1 illustrates the schematics of all three measures of polarization. It must be noted that if polarization decreases due to the intervention proposed in Eq. (4), ideally, all three quantifiers must decrease.

We also define f_{ext} as the fraction of agents with conviction $|x| > x_{th}$, where x_{th} (chosen to be five) is a positive threshold. This quantifies the prevalence of extreme opinions among the agents, which at least should not increase when we nudge the agents.

IV. RESULTS

With the intervention strategy introduced in Sec. III, we find that with sufficiently small random nudge probability p, significant depolarization can be obtained, which is evident as the opinion distributions approach toward a unimodal distribution along with the decay of all three measures of polarization. To see the effects of nudge, we perform numerical simulations of the basic model in Eq. (2) using the interaction probability given in Eq. (3) and the intervention model in Eq. (4). The simulations are performed with $N = 5000$ agents for 1000 time steps with $dt = 0.01$. At initial time, x_i is uniformly chosen from a small interval, i.e., $x_i \in [-1, 1]$ for $i = 1, 2, \ldots, N$. The model parameters are chosen to be $\alpha = 3$, $\beta = 3$, $K = 3$, $m = 10$, $\gamma = 2.1$, $\varepsilon = 0.01$, and $r = 0.5$ for all the simulations unless mentioned otherwise. The parameters chosen for the simulations lead to a polarized state in the original model without intervention.

In Fig. 2, we show the contrast between the trajectories of individual opinions and the opinion distribution with and without the application of a nudge. In the absence of nudge ($p = 0$), the simulation results in Fig. 2(a) show fewer trajectories with opinions $x_i \approx 0$. This leads to a bimodal distribution of opinions characteristic of a polarized state. In contrast, in Fig 2(b), a small nudge with a probability of $p = 0.01$ is applied, and we find significantly more trajectories with moderate opinions. This, effectively, is seen to lead to an absence of polarization, and is evident from the unimodal opinion distribution. The magnifications of the region around $x_i = 0$ and its distribution (shown in Fig. 2) reveal a clear distinction between these two scenarios.

To examine the effect of network nudge, we analyze the underlying time-averaged structures of the temporal interactions network. Without nudge, the interaction network has two distinct clusters; most of the connections are among positive opinionated agents or negative opinionated agents. There exist very few connections between these two groups other than for the agents with extreme opinions. This is expected since the agents with extreme opinions are also those who tend to be more active on social networks fora; hence, on average, they form more connections. This enables them to be relatively more connected to the agents with opposing opinions. These results are visually depicted in Fig. 3 as two snapshots of evolving network diagrams. If $p = 0$, no nudge is applied. In this case, as Fig 3(b) shows, a polarized network, made up of two distinct blue- and red-colored clusters, is formed. Blue color corresponds to nodes with $x > 0$, and red color to $x < 0$. The opinion distribution shown in Fig. 3(a) confirms the existence of polarization.

However, when a nudge is applied, even for the case when the nudge probability is as small as $p = 0.01$, we find the network to be well mixed (large blue and red clusters have disappeared) [Fig. 3(e)], and this leads to a significantly depolarized state indicated by the approximate unimodality of the opinion distribution as shown in Fig. 3(d). The term echo chamber describes a situation where the beliefs or opinions of people are reinforced by interactions among a closed group of people who hold similar opinions. In recent years, this has been widely discussed in the context of online communities [21–24]. However, some studies appear to suggest that the effects of echo chambers are over estimated [42]. To infer the presence of echo chamber-type effects, we calculate the average opinion of the nearest neighbors (NN) of each agent

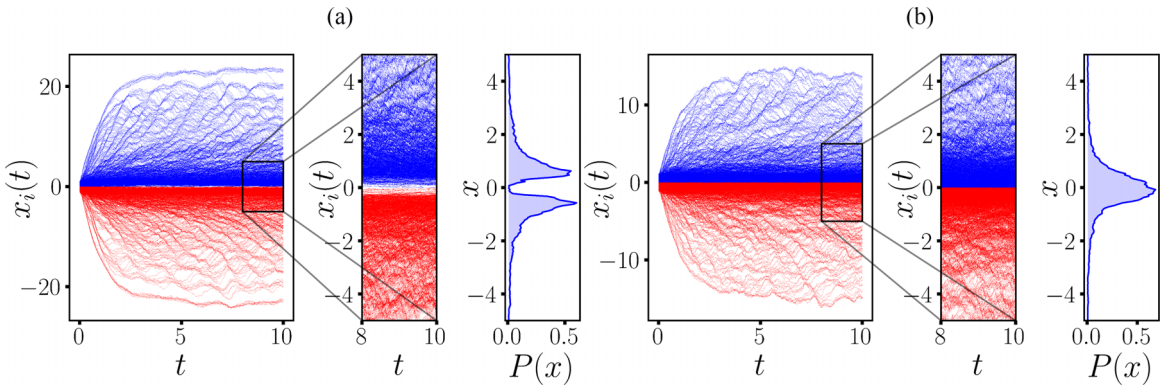

FIG. 2. Emergent polarized (and depolarized) states in the presence (and absence) of the nudge factor. The simulations are performed with 10 000 agents, and parameters are set to promote polarization. (a) The agents are not nudged. Hence the polarized state emerges. A magnification of the region around $x = 0$ reveals the absence of trajectories there, and the corresponding distribution shows a bimodal distribution with a near-zero density close to $x \approx 0$. (b) Network nudge is introduced with probability $p = 0.01$, and we find a significant depolarization. Opinion trajectories tend to crowd around $x = 0$, and the opinion distribution approaches an approximate unimodal and almost-symmetric distribution about $x = 0$.

[24,25]. This is denoted by

$$\langle x_{NN} \rangle = k_i^{-1} \sum_j a_{ij} x_j, \quad \text{and} \quad k_i = \sum_j a_{ij}, \tag{5}$$

where a_{ij} is the temporally aggregated (over the last 100 time steps) adjacency matrix. When a nudge is not applied ($p = 0$), a colored heatmap of x and $\langle x_{NN} \rangle$ in Fig. 3(c) reveals two disjoint hot spots corresponding to the two distinct echo chambers. A strong bimodality is observed in the marginal distributions. Now, when we apply a nudge with probability $p = 0.01$, we can observe only one hot spot indicating the existence of only one closed group [Fig. 3(f)]. All the agents are inside this closed group, and the echo chamber effect is largely diluted or nonexistent. We did not find perfect unimodality in the marginal distribution of x, which can be attributed to the fact that different realizations can lead to either of these three distributions: (a) slight bimodal distribution with significant reduction in all three polarization parameters, (b) unimodal distribution with a slight skew toward positive opinions, and (c) similar distribution with a skew toward negative opinions. As the heat maps and the marginal distributions are created from data averaged over 200 realizations,

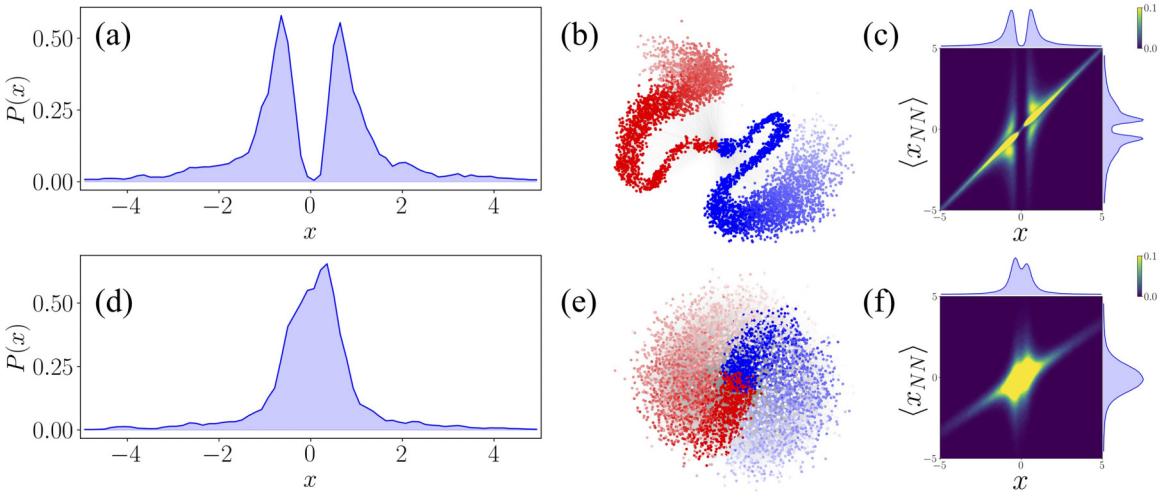

FIG. 3. Effect of the nudge on the opinion distribution, the structure of social interactions networks, and the signature of echo chambers. The networks are averaged over the last 100 time steps of simulation and are drawn using the `draw` function in `networkx` [40]. Nodes with blue color correspond to agents with positive opinions, and red corresponds to agents with negative opinions. The saturation of the color is mapped to the conviction of the agents; high saturation corresponds to a high level of conviction, and vice versa. The opinion of an agent x and the mean opinion of its nearest neighbors $\langle x_{NN} \rangle$ is averaged over 200 realizations to generate the heatmap to indicate the presence of echo chambers [see Eq. (5)]. The marginal distributions are shown in the corresponding axes. (a) For $p = 0$, i.e., without a nudge, the distribution is polarized, and the network has two distinct clusters (b), one formed by the agents with positive opinions and the other by the agents with negative opinions. (c) The presence of two distinct lobes in the heatmap indicates the echo chamber effect. (d) For $p = 0.01$, we observe an opinion distribution with a single peak, and the social interactions network is now well mixed (e). A depolarization state is reached. (f) A single lobe in the heatmap confirms the weakening of the echo chamber effect.

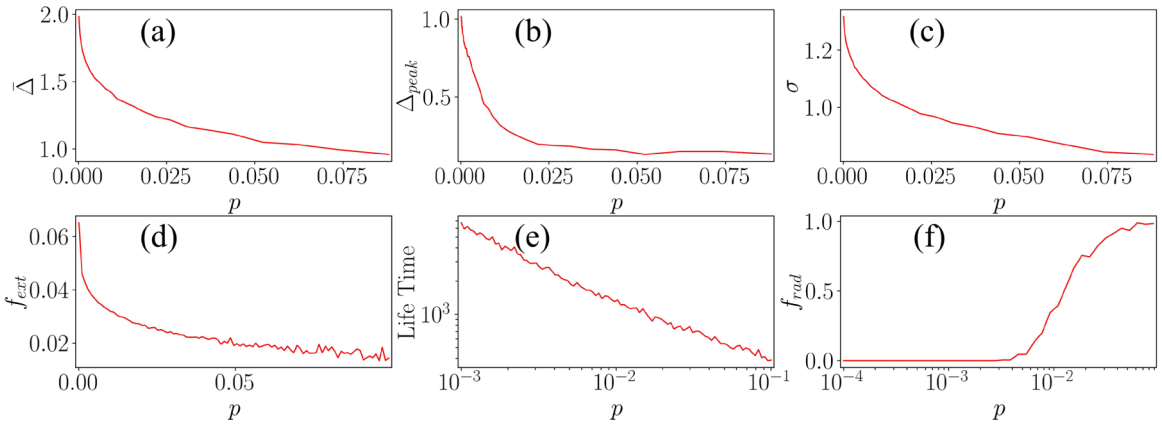

FIG. 4. Three measures of polarization, (a) $\bar{\Delta}$, (b) Δ_{peak}, (c) σ, and the fraction f_{ext} of agents with extreme opinions (d), as a function of nudge strength p. All four parameters are averaged over the last 100 time steps. The simulations were repeated 200 times, and only nonradicalized realizations were considered for ensemble averaging. The average lifetime until the whole population moves toward radicalization as a function of p is shown in panel (e). Panel (f) shows the fraction of simulations that lead to radicalization for different nudge strengths p.

all the above factors contribute to the slight bimodality in the marginal distribution of x. Nevertheless, the marginal distribution corresponds to a significant reduction in polarization and echo chambers.

V. OPTIMIZING THE NUDGE: POLARIZATION VERSUS RADICALIZATION

To obtain a global picture of how depolarization sets in as a function of nudge probability p, we plot the three measures of polarization as a function of p. All three measures, $\bar{\Delta}$, Δ_{peak}, and σ, have been computed from the simulation results. The results shown represent an average over the last 100 time steps of the simulations and averaged over 200 realizations. In Fig. 4, we observe that all three measures of polarization decrease as the strength of the nudge p increases. In particular, $\bar{\Delta}$ and σ are found to decrease as a stretched exponential function $\exp(-p^{\gamma})$, and the stretching factor γ is determined through regression to be approximately 0.3. A recent work studying the depolarization of echo chambers [41] considered adding an effective noise term dependent on a random sample of opinions to Eq. (2). While this approach succeeds in making the opinion distribution unimodal, it increases the width of the distribution significantly, which as a consequence, corresponds to an increase in extreme opinions. In contrast, the framework of nudging the mechanism of forming social connections in online interactions works well in decreasing width of the opinion distribution [Fig. 4(c)] as well as extreme opinions [Fig 4(d)] and also suggests direct algorithmic interventions for recommender systems.

In the original model, the authors found the polarized state to be metastable and showed that with an increased value of β, the lifetime of the state has a faster than exponential growth. Our intervention adds more randomness to the system and increases statistical fluctuations. Hence, for large p, we observe a drastic decrease in the average lifetime of the polarized and depolarized states. An approximate straight line in the log-log plot indicates the lifetime of polarized or depolarized states decreases as a power law as nudge strength p is increased [see Fig. 4(e)]. Figure 4(f) also captures the same effect as we see that radicalization is either nonexistent or a rarity for $p < 10^{-2}$, but it increases quickly and becomes the norm for $p > 10^{-2}$.

In many situations, radicalization is as much undesirable as polarization. Hence, to solve the issue of radicalization at a high value of nudge probability, rather than nudging all the people in the population, at each time step of the simulation, we randomly selected a fraction f of the population and nudged them. We define a simple linear utility function $U(\bar{\Delta}, f_{\text{rad}}) = \tilde{\bar{\Delta}} + f_{\text{rad}}$, where $\tilde{\bar{\Delta}}$ is $\bar{\Delta}$, linearly scaled to be between zero and one, and f_{rad} is the fraction of radicalized simulations. The structure of the utility function is the same for the other two measures of polarization. Figure 5 depicts the heat map of the utility functions corresponding to the three utility functions. The optimal population fraction and nudge probability is numerically found to follow the curve $p \cdot f^{A} = B$, where A and B are constants.

VI. ROBUSTNESS OF THE FRAMEWORK

To ensure the robustness of our intervention framework, we applied network nudge to another recent model of opinion dynamics, namely the social compass model [43,44], which, together with homophily, exhibits the effect of echo chambers. The original model describes the dynamics of opinions on two interdependent topics in polar coordinates. We reinterpret the polar angle in the original model as the opinion on a single topic to adapt the model to our framework. The dynamics of this modified model is governed by the following N coupled differential equation:

$$\dot{x}_i(t) = |x_i| \sin\left(x_i^0 - x_i\right) + K\left(\sum_{j=1}^{N} A_{ij}(t) \sin\left(x_i - x_j\right)\right). \quad (6)$$

In contrast to the original model [43], the variable x_i is chosen to be the opinions of the people on a single topic, and the temporal adjacency matrix is formed according to homophily probability 3. x_i^0 is the initial opinion of agent i, and all the other variables and parameters have the same meaning as in

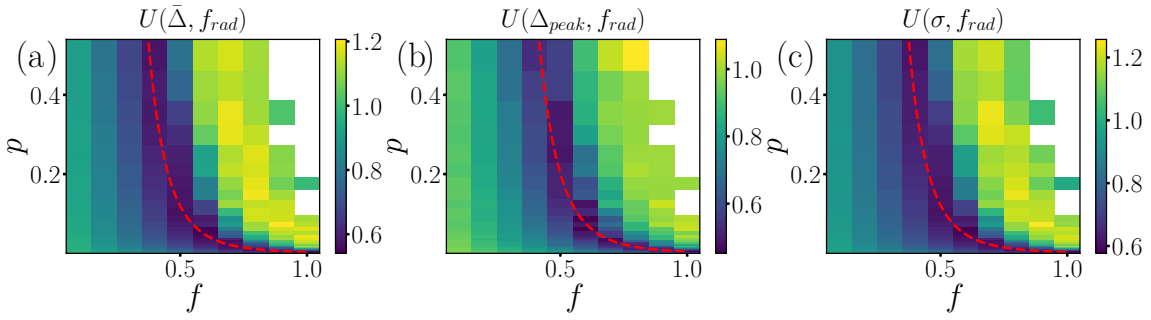

FIG. 5. The heat map of the utility as a function of nudge strength and nudged population fraction. Panels (a), (b), and (c) correspond to the corresponding utility of $\bar{\Delta}$, Δ_{peak}, and σ, respectively. The red dashed curve, which is found to follow the curve $p \cdot f^A = B$, (A, B = constants), denotes the optimal values of population fraction and nudge strength.

the previous model 2. In Fig. 6, we show that when the social interaction and the homophily factor are high enough ($K = 4$, $\beta = 4$), many echo chambers are formed, which is clear from the trajectories of the opinion as well as from the multiple communities seen in the aggregated network [Figs. 6(a) and 6(b)]. But when we introduce a slight nudge with $p = 0.002$, the effect of echo chambers is reduced drastically. The opinion trajectories seem to converge to a moderate value, and the interaction network is well connected without any obvious segregated communities Figs. 6(c) and 6(d).

VII. DISCUSSION

The widespread use of the internet, and consequently, social media platforms, have drastically altered the way humans consume, interact with, and exchange information. Polarization and the formation of echo chambers have been shown to negatively impact constructive discussions and debates—two fundamental pillars of a healthy democracy. Building on the recent advances in the modeling of opinion dynamics in social networks, in this work, we study the possibility of depolarizing a population using a stochastic nudge.

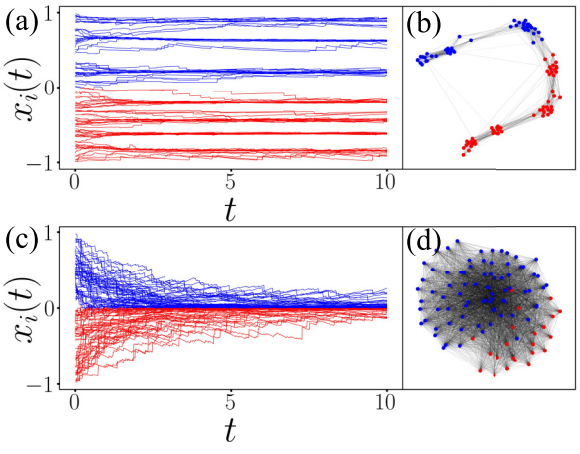

FIG. 6. The effect of nudge in the opinion dynamics model, governed by Eq. (6). Panels (a) and (c) show the trajectories of opinions in the absence and presence of network nudge, respectively. Panels (b) and (d) show the corresponding interaction network structure. Clearly, we see the presence of echo chambers in the absence of a network nudge, and the effect decreases when a slight nudge is applied.

Our results suggest that a small number of randomized interactions, which are otherwise dominated by homophily driven mechanisms, can lead to a significant reduction in polarization. This reduction was quantitatively captured by three different measures of polarization. While we show that minimal nudges can burst echo chambers and lead to socially desirable distributions of opinions, increasing the strength of this nudge can result in radicalization. Given this sensitivity on the nudge strength, we show that a possible resolution is obtained if, instead of nudging each agent, only a fraction f of the agents are nudged. We highlight that this interplay of the nudge strength p and the fraction f of nudged individuals leads to an interesting optimization problem. This optimization can help inform the fraction of individuals to be nudged for a fixed nudge strength for optimal depolarization.

We believe that the strongest case for the application of such randomized nudges can be made to recommendation systems. While ubiquitous, recommender algorithms are optimized for increasing engagement [45], which we now know can come at the cost of creating echo chambers [46], increase in the representation of extreme ideologies [47], and even the tampering of users' preferences [48]. In such settings, the randomized nudges can be potentially operationalized as the poisoning of a viewer's watch history with a limited amount of random content, uncorrelated with the viewer's preferences [49]. While there are several ethical and legal considerations that must be accounted for before implementing any such interventions, it certainly opens up several interesting avenues for future research to build on. Noninvasive interventions may be important to reduce the detrimental effects of polarization. However, an important first step is to build reliable tools to quantify polarization from data [50], which in itself constitutes an intriguing direction for future research.

ACKNOWLEDGMENTS

R.P. and A.K. gratefully acknowledge the Prime Minister's Research Fellowship of the Government of India for financial support. M.S.S. acknowledges the support of MATRICS Grant MTR/2019/001111 from SERB, Government of India. The authors acknowledge the National Supercomputing Mission for the use of PARAM Brahma at IISER Pune.

[1] S. C. McGregor, Social media as public opinion: How journalists use social media to represent public opinion, Journalism **20**, 1070 (2019).

[2] M. H. DeGroot, Reaching a consensus, J. Am. Stat. Assoc. **69**, 118 (1974).

[3] R. A. Holley and T. M. Liggett, Ergodic theorems for weakly interacting infinite systems and the voter model, Ann. Probab. **3**, 643 (1975).

[4] S. Redner, Reality-inspired voter models: A mini-review, C. R. Phys. **20**, 275 (2019).

[5] K. Sznajd-Weron and J. Sznajd, Opinion evolution in closed community, Int. J. Mod. Phys. C **11**, 1157 (2000).

[6] K. Sznajd-Weron, J. Sznajd, and T. Weron, A review on the sznajd model-20 years after, Physica A **565**, 125537 (2021).

[7] C. G. Lord, L. Ross, and M. R. Lepper, Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. J. Pers. Soc. Psychol. **37**, 2098 (1979).

[8] P. DiMaggio, J. Evans, and B. Bryson, Have american's social attitudes become more polarized? Am. J. Sociol. **102**, 690 (1996).

[9] D. Baldassarri and A. Gelman, Partisans without constraint: Political polarization and trends in american public opinion, Am. J. Sociol. **114**, 408 (2008).

[10] R. Axelrod, The dissemination of culture: A model with local convergence and global polarization, J. Conflict Resol. **41**, 203 (1997).

[11] S. Galam, Y. Gefen, and Y. Shapir, Sociophysics: A new approach of sociological collective behaviour. i. mean-behaviour description of a strike, J. Math. Sociol. **9**, 1 (1982).

[12] S. Galam and S. Moscovici, Towards a theory of collective phenomena: Consensus and attitude changes in groups, Eur. J. Soc. Psychol. **21**, 49 (1991).

[13] S. Galam, Minority opinion spreading in random geometry, Eur. Phys. J. B **25**, 403 (2002).

[14] S. Galam, Sociophysics: A Physicist's Modeling of Psycho-political Phenomena, *Understanding Complex Systems* (Springer, New York, 2012).

[15] S. Galam and F. Jacobs, The role of inflexible minorities in the breaking of democratic opinion dynamics, Physica A **381**, 366 (2007).

[16] S. Galam, Stubbornness as an unfortunate key to win a public debate: An illustration from sociophysics, Mind & Society **15**, 117 (2016).

[17] S. Galam, Collective beliefs versus individual inflexibility: The unavoidable biases of a public debate, Physica A **390**, 3036 (2011).

[18] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, Mixing beliefs among interacting agents, Adv. Complex Syst. **03**, 87 (2000).

[19] R. Hegselmann and U. Krause, Opinion dynamics and bounded confidence: Models, analysis, and simulation, J. Art. Soc. Soc. Simul. (JASSS) **5** (2002).

[20] R. K. Garrett, Echo chambers online?: Politically motivated selective exposure among internet news users, J. Comput.-Mediat. Comm. **14**, 265 (2009).

[21] M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi, Echo chambers: Emotional contagion and group polarization on facebook, Sci. Rep. **6**, 37825 (2016).

[22] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, Quantifying echo chamber effects in information spreading over political communication networks, EPJ Data Science **8**, 35 (2019).

[23] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship, in *Proceedings of the 2018 World Wide Web Conference* (ACM Digital Library, New York, 2018) pp. 913–922.

[24] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, The echo chamber effect on social media, Proc. Natl. Acad. Sci. **118**, e2023301118 (2021).

[25] F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini, Modeling Echo Chambers and Polarization Dynamics in Social Networks, Phys. Rev. Lett. **124**, 048301 (2020).

[26] F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini, Emergence of Polarized Ideological Opinions in Multidimensional Topic Spaces, Phys. Rev. X **11**, 011012 (2021).

[27] F. P. Santos, Y. Lelkes, and S. A. Levin, Link recommendation algorithms and dynamics of polarization in online social networks, Proc. Natl. Acad. Sci. **118**, e2102141118 (2021).

[28] K. Sasahara, W. Chen, H. Peng, G. L. Ciampaglia, A. Flammini, and F. Menczer, Social influence and unfollowing accelerate the emergence of echo chambers, J. Comput. Soc. Sci. **4**, 381 (2021).

[29] M. McPherson, L. Smith-Lovin, and J. M. Cook, Birds of a feather: Homophily in social networks, Annu. Rev. Sociol. **27**, 415 (2001).

[30] A. Bessi, F. Petroni, M. D. Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, and W. Quattrociocchi, Homophily and polarization in the age of misinformation, Eur. Phys. J.: Spec. Top. **225**, 2047 (2016).

[31] V. V. Vasconcelos, S. M. Constantino, A. Dannenberg, M. Lumkowsky, E. Weber, and S. Levin, Segregation and clustering of preferences erode socially beneficial coordination, Proc. Natl. Acad. Sci. **118**, e2102153118 (2021).

[32] P. Törnberg, Echo chambers and viral misinformation: Modeling fake news as complex contagion, PLoS One **13**, e0203958 (2018).

[33] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, The spreading of misinformation online, Proc. Natl. Acad. Sci. **113**, 554 (2016).

[34] D. G. Myers and H. Lamm, The group polarization phenomenon. Psychological Bulletin **83**, 602 (1976).

[35] D. J. Isenberg, Group polarization: A critical review and meta-analysis. J. Pers. Soc. Psychol. **50**, 1141 (1986).

[36] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, Activity driven modeling of time varying networks, Sci. Rep. **2**, 469 (2012).

[37] M. Starnini and R. Pastor-Satorras, Topological properties of a time-integrated activity-driven network, Phys. Rev. E **87**, 062807 (2013).

[38] A. Moinet, M. Starnini, and R. Pastor-Satorras, Burstiness and Aging in Social Temporal Networks, Phys. Rev. Lett. **114**, 108701 (2015).

[39] S. Liu, N. Perra, M. Karsai, and A. Vespignani, Controlling Contagion Processes in Activity Driven Networks, Phys. Rev. Lett. **112**, 118702 (2014).

[40] A. Hagberg, P. Swart, and D. S Chult, Exploring network structure, dynamics, and function using NetworkX, Tech. Rep. No. LA-UR-08-05495, Los Alamos National Lab. (LANL), Los Alamos, NM, United States (2008).

[41] C. B. Currin, S. V. Vera, and A. Khaledi-Nasab, Depolarization of echo chambers by random dynamical nudge, Sci. Rep. **12**, 9234 (2022).

[42] E. Dubois and G. Blank, The echo chamber is overstated: The moderating effect of political interest and diverse media, Inf. Commun. Soc. **21**, 729 (2018).

[43] J. Ojer, M. Starnini, and R. Pastor-Satorras, Modeling Explosive Opinion Depolarization in Interdependent Topics, Phys. Rev. Lett. **130**, 207401 (2023).

[44] J. Ojer, M. Starnini, and R. Pastor-Satorras, Vanishing threshold in depolarization of correlated opinions on social networks, arXiv:2306.01329.

[45] S. Milano, M. Taddeo, and L. Floridi, Recommender systems and their ethical challenges, AI Soc. **35**, 957 (2020).

[46] E. Noordeh, R. Levin, R. Jiang, and H. Shadmany, Echo chambers in collaborative filtering based recommendation systems, arXiv:2011.03890.

[47] J. Whittaker, S. Looney, A. Reed, and F. Votta, Recommender systems and the amplification of extremist content, Internet Policy **10**, 1 (2021).

[48] C. Evans and A. Kasirzadeh, User tampering in reinforcement learning recommender systems, arXiv:2109.04083.

[49] M. Haroon, A. Chhabra, X. Liu, P. Mohapatra, Z. Shafiq, and M. Wojcieszak, Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations, arXiv:2203.10666.

[50] M. Hohmann, K. Devriendt, and M. Coscia, Quantifying ideological polarization on a network using generalized euclidean distance, Sci. Adv. **9**, eabq2044 (2023).

Universal Statistics of Competition in Democratic Elections

Ritam Pal,* Aanjaneya Kumar,† and M. S. Santhanam‡
Department of Physics, Indian Institute of Science Education and Research, Pune 411008, India

(Received 26 January 2024; accepted 4 November 2024; published 9 January 2025)

Elections for public offices in democratic nations are large-scale examples of collective decision-making. As a complex system with a multitude of interactions among agents, we can anticipate that universal macroscopic patterns could emerge independent of microscopic details. Despite the availability of empirical election data, such universality, valid at all scales, countries, and elections, has not yet been observed. In this Letter, we propose a parameter-free voting model and analytically show that the distribution of the victory margin is driven by that of the voter turnout, and a scaled measure depending on margin and turnout leads to a robust universality. This is demonstrated using empirical election data from 34 countries, spanning multiple decades and electoral scales. The deviations from the model predictions and universality indicate possible electoral malpractices. We argue that this universality is a stylized fact indicating the competitive nature of electoral outcomes.

DOI: 10.1103/PhysRevLett.134.017401

One of the cornerstones of democratic societies is that governance must be based on an expression of the collective will of the citizens. The institution of elections is central to the operational success of this system. Elections to public offices are the best-documented instances of collective decision-making by humans, whose outcome is determined by multiple agents interacting over a range of spatial and temporal scales. These features make elections an interesting test bed for statistical physics whose key lesson is that a multitude of complex interactions between microscopic units of a system can manifest into robust, *universal* behavior at a macroscopic level [1–13]. A collection of gas molecules or spins are examples that display such emergent macroscopic features [14], and so are complex processes such as earthquakes [15,16] and financial markets [17]. In the context of elections, such universal behaviors serve to distill the complexities of electoral dynamics into understandable and predictive frameworks and safeguard its integrity.

Unsurprisingly, the possibility of universality in elections attracts significant research attention [18–24]. Several works have studied and proposed models for (i) the distribution $q(\sigma)$ of the fraction of votes σ obtained by candidates (or the vote share) and (ii) distribution $g(\tau)$ of voter turnout τ. While σ is indicative of popularity, τ indicates the scale of the election. Though some universality has been observed in $q(\sigma)$ or $g(\tau)$ within a single country [18–20] or in countries with similar election protocols [19,23], deviations from claimed universalities have also been reported [23,25–28] due to variations in the size (scale) of electoral districts and weak party associations. Though voting patterns tend to display spatial correlations [29–32], it is not known to be universal. Despite the availability of enormous election data and persistent attempts, a robust and universal emergent behavior, valid across different scales and countries with vastly different election protocols, is yet to be demonstrated.

In this Letter, using extensive election data [33–36] from 34 countries (from six continents) spanning multiple decades and electorate scales, we demonstrate universality through analysis of the margin of victory and turnout data in democratic elections. The *margin of victory* (or simply the *margin*) is a key indicator of competition in elections and a proxy for the healthy functioning of democracies. While the turnout data has been studied in various settings, margins have never been considered in the context of universality. We propose a random voting model (RVM) and demonstrate that the turnout distribution drives the distribution of scaled margin; i.e., the model predicts the scaled margin distribution with only the turnout distribution as the input. We analytically derive the distribution of scaled margin-to-turnout ratio in the RVM and show that it exhibits universal characteristics independent of the turnout distribution. Remarkably, we find that empirical election data across 32 countries shows excellent agreement with the analytical results, establishing a robust universality. We demonstrate its utility as a novel statistical indicator for flagging electoral malpractices [37,38].

A template of a basic electoral process is as follows. At each electoral unit, candidates compete against each other to win the votes of the electorate, who can cast their vote in

*Contact author: ritam.pal@students.iiserpune.ac.in
†Contact author: aanjaneya@santafe.edu
Present address: Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA.
‡Contact author: santh@iiserpune.ac.in

favor of only one of the candidates. The candidate securing the largest number of polled votes is declared the winner. This represents the core process in many electoral systems. It is the standard first-past-the-post system followed in many countries, e.g., India, the UK, and the USA. In an instant-runoff system (such as in Australia) or two-round runoffs (such as in France), the final runoff round boils down to this template. Typically, national or regional elections following this template consist of many electoral units made up of polling booths, precincts, constituencies, or counties. These units set a size scale in terms of the number of electorates—the polling booth represents the smallest scale, while a constituency (subsuming many polling booths) represents the largest scale. For our analysis, an "election" could be either a national, regional, or even a city-level electoral process encompassing N electoral units, and each unit could be a polling booth, county, or constituency.

In any such election, an informative indicator of the degree of competition and the extent of consensus is the margin. A vanishing margin signifies tight competition and a divided electorate, whereas large margins indicate a decisive mandate and overwhelming consensus in favor of one candidate. Let $c_i, i = 1, 2, \ldots, N$, denote the number of candidates contesting an election in the ith electoral unit. The winning and runner-up candidates receive, respectively, $v_{i,w}$ and $v_{i,r}$ votes such that $v_{i,w} > v_{i,r}$. The margin is given by $M_i = v_{i,w} - v_{i,r}$. If $n_i > 0$ is the size of the electorate, i.e., number of registered voters in ith unit, then $0 \leq M_i \leq n_i$. However, in practice, only a fraction of the electorate participates in voting. In such cases, the number of voters who show up to cast their vote is termed as the turnout T_i, such that $0 \leq T_i \leq n_i$, and consequently, the margin is further restricted by $0 \leq M_i \leq T_i$.

To fix our ideas, we might focus on the elections in one country, e.g., the general elections in India. Then, the object of interest would be M_i and T_i $(i = 1, 2, \ldots, N)$. To be statistically robust, the data is consolidated from many elections spread over several decades (For India, 18 elections from 1951 to 2019; See Sec. S6 of Supplemental Material [39]). This leads to the associated empirical distributions $Q(M)$ and $g(T)$, respectively, for margin and turnout. Figure 1(a) displays the distribution of raw turnout $g(T)$ at the constituency level for national elections in six countries, namely, India, the USA, South Korea, Canada, Japan, and Germany. Striking dissimilarities in $g(T)$ are visible in the shape and support of distribution for countries. For Germany, $g(T)$ has a unimodal character, while that for Canada and the USA display multiple peaks. The corresponding scaled margin $M/\langle M\rangle$ is displayed as distribution $f(M/\langle M\rangle)$ (computed from the consolidated margin data for each country) in Figs. 1(b)–1(g). While they appear to be broadly similar, certain differences are clearly noticeable. In particular, $f(M/\langle M\rangle)$ for German elections in Fig. 1(g) has a sharp cutoff, but for India and Japan in Figs. 1(b) and 1(f) the distribution has a slower decay. These observations motivate the questions of whether $f(M/\langle M\rangle)$ is related to the raw turnout distribution and can be obtained from it.

To investigate this question, we propose a random voting model (RVM) $\mathcal{V}(T)$ that takes raw turnouts $T = \{T_1, T_2, \ldots, T_N\}$ as input. This model emulates an election taking place at N electoral units (say, constituencies). At ith unit, each of the T_i voters (raw turnout at ith unit) can cast only one vote, independently and by randomly choosing one of the c_i contesting candidates. The probability that candidate j in ith unit can attract a vote is $p_{ij} = w_{ij}/\sum_k w_{ik}$, where $w_{ij} \in [0, 1]$ is a random number drawn from a uniform distribution. While this protocol provides a natural and effective choice for p_{ij}, the sensitivity of the RVM predictions on different protocols is discussed in Sec. S5 of Ref. [39]. In election data that we use, averaged over all the 34 countries, the top two or three candidates account for 79% (87%) of all votes polled. Hence, the

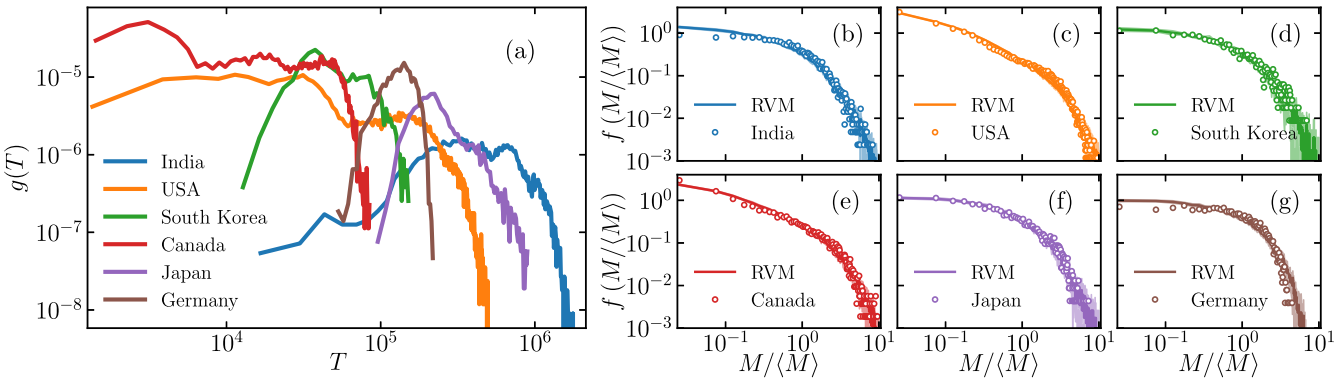

FIG. 1. (a) Turnout distribution $g(T)$ obtained from election data for different countries. Note the differences in shapes and ranges for $g(T)$. (b)–(g) Scaled margin distribution $f(M/\langle M\rangle)$ obtained from election data (open circles) and the model predictions (solid lines) display an excellent agreement. The lighter shade around the model prediction represents its variability estimated from multiple RVM realizations.

model assumes three candidates at every constituency: $c_i = 3$ for $i = 1, 2, \ldots, N$, and that all eligible voters cast their votes, implying $T_i = n_i$. By simulating this model, margin M_i is obtained for ith electoral unit and $\langle M \rangle = (1/N)\sum_{i=1}^{N} M_i$ is the associated sample mean. For a detailed description of the model, see Sec. S1 of Ref. [39].

The model predictions depend exclusively on the actual turnout distribution and no free parameters to be tuned. As illustrated in Figs. 1(b)–1(g), the scaled margin distributions predicted by this model (solid lines) show a remarkable agreement with those computed from empirical margin data from real elections. Notably, RVM faithfully captures disparate decay features in $f(M/\langle M \rangle)$ for India, the USA, South Korea, Canada, Japan, and Germany (for 28 other countries, see Sec. S7 of Ref. [39]). This suggests that the raw turnout data carries intrinsic information about the margin distribution. RVM effectively leverages this information embedded in the turnout distribution to predict the scaled margin distribution.

Next, we show that these results are independent of the number of voters or size of electoral units. In large countries, depending on the size of the electoral unit, the typical turnout can differ by several orders of magnitude. For example, in India, polling booths have a typical electoral size $\sim 10^3$, whereas, at the parliamentary constituency level, it is about 10^6. Further, the shapes of $g(T)$ are also vastly different at different scales. Figure 2(a) captures the striking differences in range and shape of $g(T)$ for India, the US, and Canada at two different scales. Quite remarkably, despite these vast differences in the scale, the same RVM $\mathcal{V}(T)$, without any parameter adjustments, accurately predicts the scaled margin distribution. Figures 2(b)–2(d) show the empirical distribution of scaled margins (in national elections) at the constituency-level scale, and Figs. 2(e)–2(g) shows the same at the scale of polling booths (county for USA). The margin distribution computed from the model is in agreement with the empirical distribution at both scales. Theoretical analysis in the limit $T \gg 1$ (see Sec. S3 of Ref. [39]) shows that the tail of $g(T)$ dictates the tail of the $f(M/\langle M \rangle)$. This is confirmed by the RVM simulations (see Sec. S4 of Ref. [39]). In particular, this is evident for the USA, where the county-level turnout distribution shows a heavy-tailed decay, which is reflected in the corresponding scaled margin distribution [Fig. 2(f)]. The faster decay at congressional district level distribution [Fig. 2(c)] is also predicted by RVM. For Canada too, the empirical scaled margin distributions are noticeably different at two different scales. Yet, the differences are well captured by the RVM simulations shown as dashed and solid lines in Figs. 2(b)–2(g). Taken together, these results show that the scaled margin distribution depends on the raw turnout distribution, and RVM captures this relation across various countries and at all scales. Then, a relevant quantity of interest would be the ratio $\mu = \frac{M}{T}$, to be called the specific margin, with $0 < \mu < 1$. This is a turnout-independent measure of electoral competitiveness and does not depend on the size of the electorate.

To obtain analytical insight, we consider elections with three candidates in the limit of large turnout ($T \gg 1$). The votes received by jth candidate can be approximated as $v_j \approx p_j T$, and the margin as $M \approx (p_{(3)} - p_{(2)})T$, where $p_{(k)}$ denotes kth order statistics [41] of the probabilities assigned to the candidates. Evidently, in this limit, $\mu \approx p_{(3)} - p_{(2)}$ and its distribution has no explicit dependence on T. With this insight, we obtain the distribution of specific margins as [39]

$$P(\mu) = \frac{(1-\mu)(5+7\mu)}{(1+\mu)^2(1+2\mu)^2}. \tag{1}$$

Thus, the distribution $F(x)$ of the scaled specific margin $x = \mu/\langle \mu \rangle$, can be expressed as

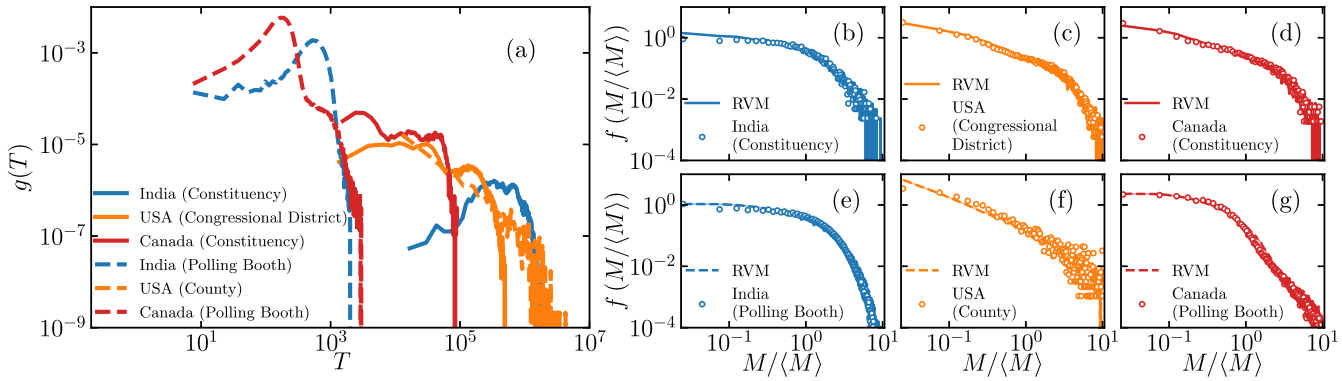

FIG. 2. The turnout distribution $g(T)$ and scaled margin distribution $f(M/\langle M \rangle)$ for India (blue), the USA (orange), and Canada (red), at two widely different scales, i.e., size of electoral units. (a) $g(T)$ at two different scales for each country. The dashed line is for smaller scales (polling booth for India and Canada, county for the USA), while the solid line represents a larger scale (constituency for India and Canada, congressional district for the USA). (b)–(g) $f(M/\langle M \rangle)$ from election data (open circles) and as predicted by the RVM (line). Despite the differences in scale and shape of $g(T)$, the empirical $f(M/\langle M \rangle)$ is well described by the RVM. The lighter shade around the model prediction represents variability estimated from multiple RVM realizations.

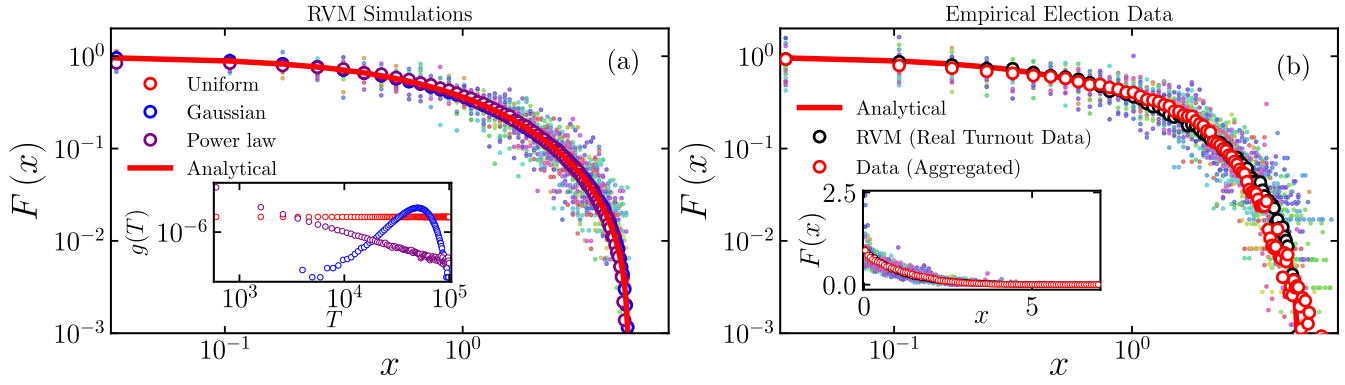

FIG. 3. (a) $F(x)$ predicted by RVM for three different turnout distributions $g(T)$ (see inset). The open circles are obtained from RVM simulations with $N = 10^6$, while the solid colored circles are generated from RVM simulation with N identical to empirical election data. The red line corresponds to $F(x)$ in Eq. (2). (b) The empirical distribution of $x = \mu/\langle\mu\rangle$ from election data of 32 countries (excluding Ethiopia and Belarus). Each color indicates a specific country for which the empirical election data is consolidated over several elections. The average of these empirical distributions (red open circles) closely follows the analytical curve (red line) and the averaged RVM predictions for each country (black open circles). The inset depicts the distributions on a linear scale.

$$F(x) = \langle\mu\rangle P(x\langle\mu\rangle), \tag{2}$$

with $\langle\mu\rangle = \frac{1}{2} + \ln(9\sqrt[4]{3}/16)$. Figure 3(a) demonstrates that $F(x)$, computed from RVM simulations with vastly different turnout distributions $g(T)$, does not depend on the detailed structure of $g(T)$ and is in agreement with the analytical prediction in Eq. (2).

The RVM simulations are performed with 10^6 electoral units (for simulation details, see Sec. S4 of Ref. [39]) using $g(T)$ corresponding to power law, Gaussian, and uniform distributions [inset of Fig. 3(a)]. The simulated distributions [open circles in Fig. 3(a)], for the three cases of $g(T)$, collapse on the analytical prediction $F(x)$ (red line).

Bolstered by the ability of RVM to capture the statistics of real elections in Figs. 1 and 2, we examine if this universality prediction in Eq. (2) holds good for the empirical election data. Indeed, as observed in Fig. 3(b), the RVM prediction (black open circles) is in excellent agreement with the averaged distributions (open red circles) obtained from all the 32 countries. The averaged empirical distribution is also consistent with the analytical universal curve $F(x)$ (red line). Further, the empirical distribution for each of the 32 countries (denoted by the solid-colored circles) closely follows the trend of $F(x)$, albeit with some fluctuations induced by the finite size of data. Similar fluctuations are evident in RVM simulations as well, seen as solid circles in Fig. 3(a), when the number of electoral units N is taken from the empirical election data (rather than fixed at 10^6) [39]. Empirical distributions shown in the inset of Fig. 3(b) demonstrate that at large x, the absolute fluctuations decrease. Thus, the universality in Fig. 3 suggests that irrespective of the finer details of election processes, the mechanism underlying the core component of any competitive election—choosing one candidate from many contenders—leads to a universal distribution for the scaled specific margin $x = \mu/\langle\mu\rangle$.

From the excellent RVM predictions of scaled margin distributions (Figs. 1 and 2) and the robustness of the universality result (Fig. 3) across different countries with a track record of fair election processes, it is reasonable to assume that any pronounced deviation from $F(x)$ in Eq. (2) might indicate a prevalence of unfair means in the election process. We search for such deviations in countries with at least 400 data points in the constituency-level election data. We find that $F(x)$ computed from data for Ethiopian election of 2010 and Belarus elections during 2004–2019 display pronounced deviations from the RVM predictions and universality as seen in Fig. 4(b). Similarly, the empirical scaled margin distribution $f(M/\langle M\rangle)$ deviates significantly from the RVM prediction [Fig. 4(a)]. This analysis in Fig. 4 strengthens the skepticism expressed in

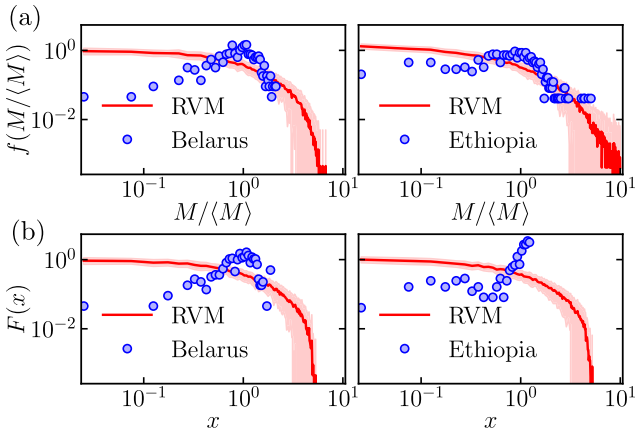

FIG. 4. The distributions (a) $f(M/\langle M\rangle)$, and (b) $F(x)$ obtained from empirical data from Belarus (2004 – 2019) and Ethiopia (2010) (blue circles). Both show significant deviation from the model predictions (red line). The light red shaded region represents the variability in RVM prediction computed from 100 realizations.

earlier studies and independent investigations about elections in Ethiopia [42] and Belarus [43–46]. Electoral malpractices take various forms, and statistical analysis is useful as a *prima facie* indicator requiring detailed scrutiny. Thus, the robust universality and RVM provide an effective toolbox to flag potentially suspicious elections. We propose that the universality in Fig. 3 should be treated as a stylized fact of elections, which all election models should be able to reproduce.

In summary, competitiveness in any election is encoded in the victory margins and turnouts. The latter also expresses people's interest in the participatory democratic process. In this work, using extensive empirical election data from 34 countries, we have obtained two significant results: (i) scaled margin distribution can be predicted from the raw election turnout alone, (ii) the scaled distribution of margin-to-turnout ratio μ has a universal form for all elections independent of country, regions, turnouts and the scale of elections. A parameter-free model introduced in this work faithfully reproduces all these features observed in empirical election data and has been analytically solved to demonstrate universality. Both these results can be regarded as stylized facts of elections. Hence, every successful election model, irrespective of its underlying principle and mechanism, must necessarily reproduce these stylized facts to be consistent with real elections. Further, the deviations from the universal scaling function could potentially help in assessing the credibility of the election process. We demonstrate this by flagging the elections of two countries for possible electoral misconduct.

Acknowledgments—The authors gratefully acknowledge the feedback of an anonymous reviewer whose suggestions greatly improved the manuscript. R. P. and A. K. thank the Prime Minister's Research Fellowship of the Government of India for financial support. M. S. S. acknowledges the support of a MATRICS Grant from SERB, Government of India, during the early stages of this work. The authors acknowledge the National Supercomputing Mission for the use of PARAM Brahma at IISER Pune.

[1] P. W. Anderson, Science **177**, 393 (1972).
[2] S. Strogatz, S. Walker, J. M. Yeomans, C. Tarnita, E. Arcaute, M. De Domenico, O. Artime, and K.-I. Goh, Nat. Rev. Phys. **4**, 508 (2022).
[3] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).
[4] A. Jedrzejewski and K. S. Weron, C.R. Phys. **20**, 244 (2019).
[5] M. San Miguel and R. Toral, Chaos **30**, 120401 (2020), see all the papers that are part of this special issue.
[6] S. Galam, *Sociophysics: A Physicist's Modeling of Psycho-Political Phenomena* (Springer, New York, NY, 2012).
[7] S. J. Brams, *Mathematics and Democracy: Designing Better Voting and Fair-Division Procedures* (Princeton University Press, Princeton, NJ, 2008).
[8] S. Fortunato, M. Macy, and S. Redner, J. Stat. Phys. **151**, 1 (2013).
[9] J.-P. Bouchaud, J. Phys. **4**, 041001 (2023).
[10] P. Sen and B. K. Chakrabarti, *Sociophysics: An Introduction* (Oxford University Press, Oxford, 2014).
[11] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki, Phys. Rep. **687**, 1 (2017).
[12] M. Jusup, P. Holme, K. Kanazawa, M. Takayasu, I. Romić, Z. Wang, S. Geček, T. Lipić, B. Podobnik, L. Wang, W. Luo, T. Klanjšček, J. Fan, S. Boccaletti, and M. Perc, Phys. Rep. **948**, 1 (2022).
[13] S. Redner, C.R. Phys. **20**, 275 (2019).
[14] F. Reif, *Fundamentals of Statistical and Thermal Physics* (McGraw Hill, Tokyo, 1965).
[15] A. Corral, Phys. Rev. Lett. **92**, 108501 (2004).
[16] A. Corral, Phys. Rev. Lett. **97**, 178501 (2006).
[17] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).
[18] R. N. C. Filho, M. P. Almeida, J. S. Andrade, and J. E. Moreira, Phys. Rev. E **60**, 1067 (1999).
[19] S. Fortunato and C. Castellano, Phys. Rev. Lett. **99**, 138701 (2007).
[20] C. Borghesi and J.-P. Bouchaud, Eur. Phys. J. B **75**, 395 (2010).
[21] M. Mantovani, H. Ribeiro, M. Moro, S. Picoli, and R. Mendes, Europhys. Lett. **96**, 48001 (2011).
[22] E. Bokányi, Z. Szállási, and G. Vattay, PLoS One **13**, 1 (2018).
[23] A. Chatterjee, M. Mitrović, and S. Fortunato, Sci. Rep. **3**, 1049 (2013).
[24] V. Hösel, J. Müller, and A. Tellier, Palgrave Commun. **5**, 1 (2019).
[25] A. Kononovicius, Acta Phys. Pol. A **133**, 1450 (2018).
[26] A. Kononovicius, J. Stat. Mech. (2019) 103402.
[27] A. M. Calvão, N. Crokidakis, and C. Anteneodo, PLoS One **10**, 1 (2015).
[28] C. Borghesi, J.-C. Raynal, and J.-P. Bouchaud, PLoS One **7**, 1 (2012).
[29] J. Fernández-Gracia, K. Suchecki, J. J. Ramasco, M. San Miguel, and V. M. Eguíluz, Phys. Rev. Lett. **112**, 158701 (2014).
[30] D. Braha and M. A. M. de Aguiar, PLoS One **12**, 1 (2017).
[31] J. Michaud, I. H. Mäkinen, A. Szilva, and E. Frisk, Appl. Network Sci. **6**, 1 (2021).
[32] S. Mori, M. Hisakado, and K. Nakayama, Phys. Rev. E **99**, 052307 (2019).
[33] Election data of india, https://www.eci.gov.in.
[34] Constituency-level elections archive [data file and codebook], http://www.electiondataarchive.org.
[35] Election data of canada, https://www.elections.ca.
[36] M. E. Data and S. Lab, County Presidential Election Returns 2000-2020, 10.7910/DVN/VOQCHQ (2018).
[37] P. Klimek, Y. Yegorov, R. Hanel, and S. Thurner, Proc. Natl. Acad. Sci. U.S.A. **109**, 16469 (2012).
[38] R. Jimenez, M. Hidalgo, and P. Klimek, Sci. Adv. **3**, e1602363 (2017).
[39] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.134.017401 for (1) the description of RVM, (2) theoretical calculations for RVM and other related discussions, (3) data summary, and (4) figures, which includes Ref. [40].

[40] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, ninth dover printing, tenth gpo printing ed. (Dover, New York, 1964).
[41] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, A First Course in Order Statistics (Society for Industrial and Applied Mathematics, 2008).
[42] G. Brigaldino, Rev. Afr. Political Econ. **38**, 327 (2011).
[43] Report of organization for security and co-operation in europe (osce), https://www.osce.org/odihr/elections/belarus (2020).
[44] M. Frear, Electoral Stud. **33**, 350 (2014).
[45] S. Bedford, Natl. Papers **49**, 808 (2021).
[46] A. Czwołek and J. Kołodziejska, Copernicus J. Political Stud. **1**, 81 (2021), https://apcz.umk.pl/CJPS/article/view/36526.

Supplemental Material for "Universal Statistics of Competition in Democratic Elections"

Ritam Pal,[1,*] Aanjaneya Kumar,[1,†] and M. S. Santhanam[1,‡]

[1] *Department of Physics, Indian Institute of Science Education and Research, Dr. Homi Bhabha Road, Pune 411008, India.*

This Supplemental Material provides further discussion and derivations which support the findings reported in the Letter, and provides details of the models and simulations used to validate the results.

CONTENTS

S1. RANDOM VOTING MODEL: DESCRIPTION

We describe a model of elections, designated as the Random Voting Model (RVM), in which c_i number of candidates contest at i-th electoral unit with n_i electors (voters). In this model, each elector from the i-th electoral unit casts their vote for j-th candidate with a probability p_{ij}. These probabilities are assigned as follows: for each candidate, a number between 0 and 1 is drawn uniformly at random, which is assigned as an unnormalized probability weight w_{ij} to that candidate. The weights are subsequently normalized to get the probability $p_{ij}, j = 1, 2 \ldots c_i$ of receiving the vote of an elector. This can be mathematically stated as

$$w_{ij} \sim \mathcal{U}(0,1) \quad \text{and} \quad p_{ij} = \frac{w_{ij}}{\sum_k w_{ik}}, \text{ with } j = 1, 2 \ldots c_i, \tag{S1}$$

where $\mathcal{U}(0,1)$ denotes a uniformly distributed random variable in $(0,1)$.

In an election, if there are n_i electors (voters) in i-th electoral unit, each elector votes for candidate j independently with probability p_{ij}. Every voter votes exactly once. The candidate receiving the most votes $v_{i,w}$ is declared the

* ritam.pal@students.iiserpune.ac.in

† Present address: Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA; aanjaneya@santafe.edu

‡ santh@iiserpune.ac.in

winner, and the candidate securing the next largest number of votes $v_{i,r}$ is the runner-up. The *margin of victory* M_i is then defined to be the vote difference between the winner and the runner-up: *i.e.* $M_i = v_{i,w} - v_{i,w}$. The empirical election data we employ (from 34 countries) shows that the top three candidates, on average, account for nearly 87% of all votes polled in an election. Hence, as part of the model specification, we fix the number of candidates in each electoral unit to be three, i.e., $c_i = 3$ for all i.

The only input to this model is the raw turnout data, i.e., the number of voters (who actually voted) in each constituency. For the model simulation, we use the turnout data of real elections as the total number of voters in different constituencies. To understand how simulations are performed, consider this notional example: if a country has $N = 100$ constituencies and data for five such elections is available. Then, the model is simulated on 500 electoral units. The number of electors in each electoral unit is taken from the consolidated turnouts. Such a simulation of election is performed multiple times to get the average distributions for scaled margins $f(M/\langle M\rangle)$ and scaled specific margins $F(x)$.

S2. COMPUTING THE DISTRIBUTION OF SPECIFIC MARGIN $\mu = \frac{M}{T}$

As done in the previous section, we consider the case where 3 candidates are contesting in an election. The weight assigned for the j-th candidate of the i-th electoral unit is w_{ij}. These weights are drawn independently at random from a uniform distribution between 0 and 1. The corresponding probability p_{ij} of receiving votes is calculated by normalizing these weights. Hence, we have the following,

$$w_{ij} \sim \mathcal{U}(0,1) \text{ and } p_{ij} = \frac{w_{ij}}{\sum_{k=1}^{3} w_{ik}};\ \text{with } j = 1,2,3. \tag{S2}$$

For the rest of the analysis, we focus on a single (i-th) electoral unit with voter turnout T and drop the corresponding index i for brevity. Hence,

$$w_{ij} := w_j \text{ and } p_{ij} := p_j. \tag{S3}$$

For large turnout ($T \gg 1$), it is reasonable to assume the number of votes received by j-th candidate is proportional to their probability p_j, in particular, $v_j \approx p_j T$. Hence, for $T \gg 1$, the *margin* can be approximated as

$$M \approx (p_{max} - p_{2nd\,max})T, \tag{S4}$$

where p_{max} and $p_{2nd\,max}$ correspond to the largest and the second largest probabilities assigned to the candidates. For example, if the probabilities p_1, p_2, and p_3 assigned to the 3 candidates are $0.1, 0.6$, and 0.3, then $p_{max} = p_2 = 0.6$ and $p_{2nd\,max} = p_3 = 0.3$. The margin M can also be written in terms of w_j as the following:

$$\begin{aligned} M &\approx \left(\frac{w_{max}}{w_1 + w_2 + w_3} - \frac{w_{2nd\,max}}{w_1 + w_2 + w_3}\right) T, \\ &= \left(\frac{w_{(3)}}{w_{(1)} + w_{(2)} + w_{(3)}} - \frac{w_{(2)}}{w_{(1)} + w_{(2)} + w_{(3)}}\right) T, \\ &= \left(\frac{w_{(3)} - w_{(2)}}{w_{(1)} + w_{(2)} + w_{(3)}}\right) T, \end{aligned} \tag{S5}$$

where $w_{(k)}$ is the k-th order statistics [1]. Hence,

$$\frac{M}{T} \approx \frac{w_{(3)} - w_{(2)}}{w_{(1)} + w_{(2)} + w_{(3)}}. \tag{S6}$$

Consider n *iid* random variables $\{X_1, X_2 \ldots X_n\}$ drawn from a distribution $\rho(x)$. When arranged in ascending order, the random variable at the k-th spot is defined as the k-th order statistics. In particular, n-th and 1-st order statistics correspond to the maximum and minimum of those n random variables, respectively. The k-th order statistics of the random variable X is denoted by $X_{(k)}$.

The joint probability density of all the order statistics of the above-mentioned n random variables, $\mathbb{P}\left(x_{(1)}, x_{(2)}, ...x_{(n)}\right)$, defined as the probability density that the random variable $X_{(k)}$ takes the value $x_{(k)}$ for $k \in \{1, 2, \ldots, n\}$, is

$$\mathbb{P}\left(x_{(1)}, x_{(2)}, ...x_{(n)}\right) = n! \prod_{k=1}^{n} \rho\left(x_{(k)}\right). \tag{S7}$$

For our case, $n = 3$ and $\rho(x) = \mathcal{U}(0, 1)$. Hence we have,

$$\mathbb{P}\left(w_{(1)}, w_{(2)}, w_{(3)}\right) = 3! = 6; \text{ with } 0 < w_{(1)} < w_{(2)} < w_{(3)} < 1, \tag{S8}$$

and $\mathbb{P}\left(w_{(1)}, w_{(2)}, w_{(3)}\right) = 0$ otherwise, with the following normalization:

$$\int_0^1 dw_{(3)} \int_0^{w_{(3)}} dw_{(2)} \int_0^{w_{(2)}} 6 dw_{(1)} = 1. \tag{S9}$$

From the joint probability distribution of all the order statistics, we calculate the approximate probability density function of specific margin $M/T = \mu$ from Eq. (S6) as follows,

$$\begin{aligned} P(\mu) &= 6 \int_0^1 dw_{(3)} \int_0^{w_{(3)}} dw_{(2)} \int_0^{w_{(2)}} \delta\left(\mu - \frac{w_{(3)} - w_{(2)}}{w_{(1)} + w_{(2)} + w_{(3)}}\right) dw_{(1)}, \\ &= 6 \int_0^1 dw_{(3)} \int_0^{w_{(3)}} \frac{w_{(3)} - w_{(2)}}{\mu^2} \mathbb{1}_{0 < \frac{w_{(3)} - \mu w_{(3)} - (1+\mu) w_{(2)}}{\mu} < w_{(2)}} dw_{(2)}, \\ &= 6 \int_0^1 dw_{(3)} \frac{(1-\mu)(5+7\mu) w_{(3)}^2}{2(1+\mu)^2(1+2\mu)^2}. \end{aligned} \tag{S10}$$

$$\tag{S11}$$

Finally, after performing this integral, we get

$$P(\mu) = \frac{(1-\mu)(5+7\mu)}{(1+\mu)^2(1+2\mu)^2}. \tag{S12}$$

The distribution $P(\mu)$ does not depend on the turnout and is universal. Now, by a change of variable to scaled specific margin defined as $x = \mu/\langle\mu\rangle$, we obtain its distribution $F(x)$ to be

$$F(x) = \langle\mu\rangle \; P(x\langle\mu\rangle) = \frac{\langle\mu\rangle(1 - x\langle\mu\rangle)(5 + 7x\langle\mu\rangle)}{(1 + x\langle\mu\rangle)^2(1 + 2x\langle\mu\rangle)^2}, \tag{S13}$$

where $\langle\mu\rangle = \frac{1}{2} + \ln\left(\frac{9\sqrt[4]{3}}{16}\right)$.

S3. DISTRIBUTION OF MARGINS AND THEIR TAIL BEHAVIORS

In this section, we obtain the distribution of margins $Q(M)$ for arbitrary turnout distribution $g(T)$, using the specific margin distribution $P(\mu)$. From the previous section, we have

$$P(\mu) = \frac{(1-\mu)(5+7\mu)}{(1+\mu)^2(1+2\mu)^2}. \tag{S14}$$

Through a simple change of variable ($M = \mu T$) we get,

$$\mathcal{P}(M|T) = \frac{(1 - M/T)(5 + 7M/T)}{T(1 + M/T)^2(1 + 2M/T)^2}. \tag{S15}$$

For an arbitrary turnout distribution $g(T)$, we obtain the distribution of M to be,

$$Q(M) = \int_M^\infty g(T)\mathcal{P}(M|T)dT = \int_M^\infty g(T)\frac{(1-M/T)(5+7M/T)}{T(1+M/T)^2(1+2M/T)^2}dT. \tag{S16}$$

Again with $u = T/M$, the above integral transforms to,

$$Q(M) = \int_1^\infty g(Mu)\frac{u(u-1)(5u+7)}{(1+u)^2(2+u)^2}du. \tag{S17}$$

We compute $Q(M)$ for different turnout distributions $g(T)$. In particular, we take $g(T)$ to be (A) exponential, (B) power law, and (C) Gaussian distributions as they have vastly different tail behaviors.

A. Exponential Turnout Distribution

In this case $g(T) = \frac{1}{\tau}e^{-T/\tau}$, with $\tau > 0$. Hence,

$$Q(M) = \int_1^\infty \frac{1}{\tau}e^{-Mu/\tau}\frac{u(u-1)(5u+7)}{(1+u)^2(2+u)^2}du, \tag{S18}$$

or,

$$Q(M) = \frac{e^{-\frac{M}{\tau}}}{\tau^2}\left(4e^{\frac{2M}{\tau}}(\tau+M)\mathrm{Ei}\left(-\frac{2M}{\tau}\right) - 9e^{\frac{3M}{\tau}}(\tau+2M)\mathrm{Ei}\left(-\frac{3M}{\tau}\right) - 4\tau\right), \tag{S19}$$

where $\mathrm{Ei}(x) = \int_{-\infty}^x \frac{e^t}{t}dt$. At large margin limit ($M \to \infty$), the asymptotic behavior of the distribution is the following (up to the leading order of M):

$$Q(M) = \frac{\tau}{3M^2}e^{-M/\tau}. \tag{S20}$$

This suggests that in the large margin limit, both the margin and its corresponding turnout distribution have an exponential decay with the same rate.

B. Power law Turnout Distribution

In this case $g(T) = \frac{\alpha-1}{T_{min}^{1-\alpha}}T^{-\alpha}$, with $\alpha > 1$ and $T > T_{min}$. Hence we have,

$$Q(M) = \int_1^\infty \frac{\alpha-1}{T_{min}^{1-\alpha}}(Mu)^{-\alpha}\frac{u(u-1)(5u+7)}{(1+u)^2(2+u)^2}du, \tag{S21}$$

or,

$$Q(M) = C(M)\frac{\alpha-1}{T_{min}^{1-\alpha}}(M)^{-\alpha}, \tag{S22}$$

where,

$$C(M) = \begin{cases} I_1(\infty) - I_1(T_{min}/M), \text{if } M \le T_{min} & \text{(S23)} \\ I_1(\infty) - I_1(1), \text{otherwise}, & \text{(S24)} \end{cases}$$

with,

$$I_1(y) = \int \frac{y^{1-\alpha}(y-1)(5y+7)}{(1+y)^2(2+y)^2}dy, \tag{S25}$$

and,

$$I_1(y) = \begin{cases} -\frac{4}{y+1} + \frac{9}{2(y+2)} - \frac{1}{4}7\ln(y) + 4\ln(y+1) - \frac{9}{4}\ln(y+2), \text{if } \alpha = 2 & \text{(S26)} \\ \frac{y^{2-\alpha}\left(16\,{}_2F_1(2, 2-\alpha; 3-\alpha; -y) - 9\,{}_2F_1\left(2, 2-\alpha; 3-\alpha; -\frac{y}{2}\right)\right)}{4(\alpha-2)}, \text{otherwise,} & \text{(S27)} \\ & \text{(S28)} \end{cases}$$

where ${}_2F_1(a, b; c; z)$ is a hypergeometric function [2], defined as,

$${}_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!} = 1 + \frac{ab}{c}\frac{z}{1!} + \frac{a(a+1)b(b+1)}{c(c+1)}\frac{z^2}{2!} + \cdots .$$

It is evident from Eq. (S22) that for $M > T_{min}$, the margin distribution decays with a power law exponent α, exactly the same as the turnout distribution.

C. Gaussian Turnout Distribution

In this case $g(T) = C_0 e^{-(T/T_0)^2}$, with $T > 0$. Hence,

$$Q(M) = \int_1^{\infty} C_0 e^{-(Mu/T_0)^2} \frac{u(u-1)(5u+7)}{(1+u)^2(2+u)^2} du. \tag{S29}$$

At large margin limit ($M \to \infty$), the asymptotic behavior of the distribution is the following (up to the leading order of M):

$$Q(M) = \frac{C_0}{12}\left(\frac{T_0}{M}\right)^4 e^{-(M/T_0)^2}, \tag{S30}$$

and it has a Gaussian decay similar to the corresponding turnout distribution.

From the asymptotic analysis of the margin distributions for the three above-mentioned turnout distributions, we provide strong evidence that the tails of the margin distributions mimic that of the corresponding turnout distribution. For completeness, we also compute the margin distribution corresponding to a uniform turnout distribution which has a finite support (no tail behavior).

D. Uniform Turnout Distribution

In this case $g(T) = \frac{1}{b-a}$, when $T \in [a, b]$, otherwise $g(T) = 0$. Hence,

$$Q(M) = \begin{cases} \frac{1}{b-a}\int_{a/M}^{b/M} \frac{u(u-1)(5u+7)}{(1+u)^2(2+u)^2} du, \text{if } M \le a & \text{(S31)} \\ \frac{1}{b-a}\int_{1}^{b/M} \frac{u(u-1)(5u+7)}{(1+u)^2(2+u)^2} du, \text{otherwise,} & \text{(S32)} \end{cases}$$

or,

$$Q(M) = \begin{cases} \frac{1}{b-a}\left(I_2(b/M) - I_2(a/M)\right), \text{if } M \le a & \text{(S33)} \\ \frac{1}{b-a}\left(I_2(b/M) - I_2(1)\right), \text{if } a > M \ge b & \text{(S34)} \\ 0, \text{ otherwise,} & \text{(S35)} \end{cases}$$

where,

$$I_2(y) = \int \frac{y(y-1)(5y+7)}{(1+y)^2(2+y)^2} dy = -\frac{4}{y+1} + \frac{18}{y+2} - 4\ln(y+1) + 9\ln(y+2). \tag{S36}$$

S4. RVM SIMULATIONS WITH SYNTHETIC TURNOUT DISTRIBUTIONS

The RVM enables us to estimate the scaled margin distribution $f(M/\langle M\rangle)$ using only the raw turnout data, indicating that $f(M/\langle M\rangle)$ is driven by the details of the turnout distribution $g(T)$. To further quantify the effect of $g(T)$ on the scaled margin distribution $f(M/\langle M\rangle)$, we simulate elections using RVM, with turnouts drawn from vastly different synthetically generated distributions. In particular, to study the tail behaviors, we use the following four different turnout distributions:

1. **Gaussian Turnout Distribution:** $g(T) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(T-\mu)^2}{2\sigma^2}\right)$, with $\mu = 50000$, $\sigma = 10000$ and $T > 0$.
2. **Exponential Turnout Distribution:** $g(T) = \frac{1}{\tau} \exp\left(-\frac{T}{\tau}\right)$, with $\tau = 50000$.
3. **Power law Turnout Distribution:** $g(T) = \frac{\alpha-1}{T_{min}^{1-\alpha}} T^{-\alpha}$, with $\alpha = 2$ and $T_{min} = 100$ (minimum possible turnout).
4. **Uniform Turnout Distribution:** $T \sim \mathcal{U}(a, b)$, with $a = 100$ and $b = 100000$. $\mathcal{U}(a, b)$ denotes uniform distribution between the range a and b.

Each of the RVM simulations was performed on 10^6 electoral units, with turnouts (rounded down to the nearest integer) drawn from one of these three distributions. The simulation demonstrates that the tail of the margin distribution mimics the turnout distribution's tail. This is evident in Fig. S1(a), (b), and (c). The tail of the margin distribution (Fig. S1 (c)) corresponding to power law turnouts decays with the same power law exponent. In the simulation with Gaussian turnout distribution, we find the tail of the margin distribution also has a Gaussian falloff (Fig. S1 (a)). Similarly, the margin distribution corresponding to exponential turnouts has an exponential tail (Fig. S1 (b)). As the probability density function of uniform turnout distribution and corresponding margin distribution have finite supports, their tails can not be properly defined. We find a sharp cutoff in the corresponding margin distribution. The analytical (semi-analytical for Gaussian turnout) predictions for the margin distributions (shown as black lines in Fig. S1) corresponding to all four aforementioned turnout distributions are in excellent agreement with the RVM simulation. In empirical county-level election data of the United States, the heavy-tailed decay of the turnout distribution is reflected in the corresponding margin distribution (Fig. S1(e)). In Fig. S1 (f), we see a similar decay trend in both margin and turnout distribution, which correspond to congressional district-level election data of the USA. We obtain the scaled margin distribution $f(M/\langle M\rangle)$ by scaling $Q(M)$ by its mean; hence, both $Q(M)$ and $f(M/\langle M\rangle)$ have similar decay and are strongly related to the corresponding turnout distribution $g(T)$.

Simulation details of the universality result: We study the scaled specific margin distribution $F(x)$ by simulating elections using RVM for the following three turnout distributions:

1. **Gaussian Turnout Distribution:** $g(T) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(T-\mu)^2}{2\sigma^2}\right)$, with $\mu = 50000$, $\sigma = 10000$ and $T > 0$.
2. **Uniform Turnout Distribution:** $T \sim \mathcal{U}(a, b)$, with $a = 100$ and $b = 100000$. $\mathcal{U}(a, b)$ denotes uniform distribution between the range a and b.
3. **Power law Turnout Distribution:** $g(T) = \frac{\alpha-1}{T_{min}^{1-\alpha}} T^{-\alpha}$, with $\alpha = 2$ and $T_{min} = 100$ (minimum possible turnout).

Turnouts drawn from these distributions are rounded down to the nearest integers. Simulations performed with a large number of electoral units (10^6) lead to a perfect collapse in the scaled specific margin distributions $F(x)$, which is in remarkable agreement with the theoretically predicted distribution, as shown in Fig. 3(a) of the letter. When simulations are performed with realistic numbers of electoral units as found in the empirical data, the corresponding scaled distributions of μ show similar fluctuations around the universal curve, as found in the empirical distributions of individual countries. To ensure realistic statistics, the number of electoral units N chosen for each simulation is the consolidated number of electoral units for each of the 32 countries. Once the number of electoral units is fixed, we randomly choose one of the three distributions mentioned above. Further N turnouts are drawn independently from that distribution, and the RVM simulation is performed on those turnouts.

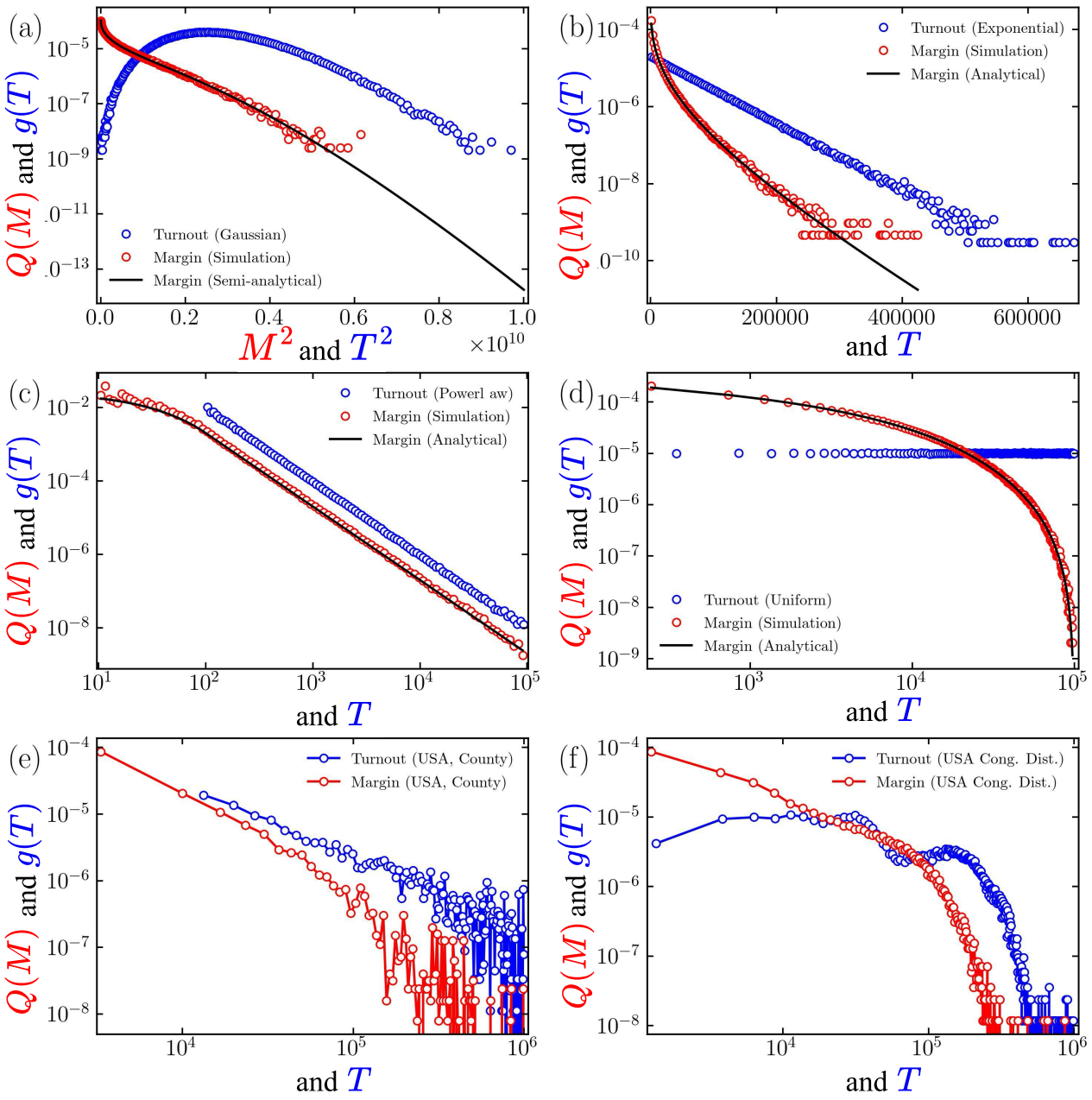

FIG. S1. The margin distribution $Q(M)$ is plotted with the corresponding turnout distribution $g(T)$ to demonstrate that the tails of both these distributions are correlated. Panels (a), (b), (c), and (d) correspond to Gaussian, exponential, power law, and uniform turnout distributions, respectively. Blue open circles denote the turnout distributions. Red open circles denote the margin distribution computed through RVM simulations. Black solid lines correspond to the margin distribution computed using Eq. S17. For exponential, power law, and uniform turnout distributions, the integration was analytically calculated, and for Gaussian turnout distribution, it was evaluated numerically. Panels (e) and (f) depict the margin and turnout distribution for the county-level and congressional district-level election data of the USA, respectively.

S5. SCALED MARGIN DISTRIBUTIONS FOR DIFFERENT p_{ij} DISTRIBUTIONS

To investigate the effect of the distribution of p_{ij} on the prediction of scaled margin distribution, we simulated RVM using the following three protocols for choosing p_{ij}.

1. **Protocol 1:** $w_{ij} \sim \mathcal{U}(0,1)$ and $p_{ij} = \frac{w_{ij}}{\sum_{k=1}^{3} w_{ik}}$; with $j = 1, 2, 3$.
2. **Protocol 2:** $w_{i1} \sim \mathcal{U}(0,1), w_{i2} \sim \mathcal{U}(0, 1 - w_{i1}), w_{i3} = 1 - w_{i1} - w_{i2}$ and $p_{ij} = \frac{w_{ij}}{\sum_{k=1}^{3} w_{ik}} = w_{ij}$; with $j = 1, 2, 3$.
3. **Protocol 3:** $w_{ij} = p_{ij} = \frac{1}{3}$, with $j = 1, 2, 3$.

In Fig. S2, we demonstrate the differences in the prediction of scaled margin distributions for synthetically generated turnout distributions when the three aforementioned protocols are used. Panel (a) shows that, for turnouts drawn from a uniform distribution, the prediction using protocols 1 and 3 are similar, while protocol 2 produces a scaled margin distribution that decays faster. In panel (b), we see similar results for turnout drawn from a Gaussian distribution. However, panel (c) depicts that, for turnouts drawn from a power law distribution, the predictions using protocols 1 and 2 are almost identical, while protocol 3 produces a vastly different scaled margin distribution.

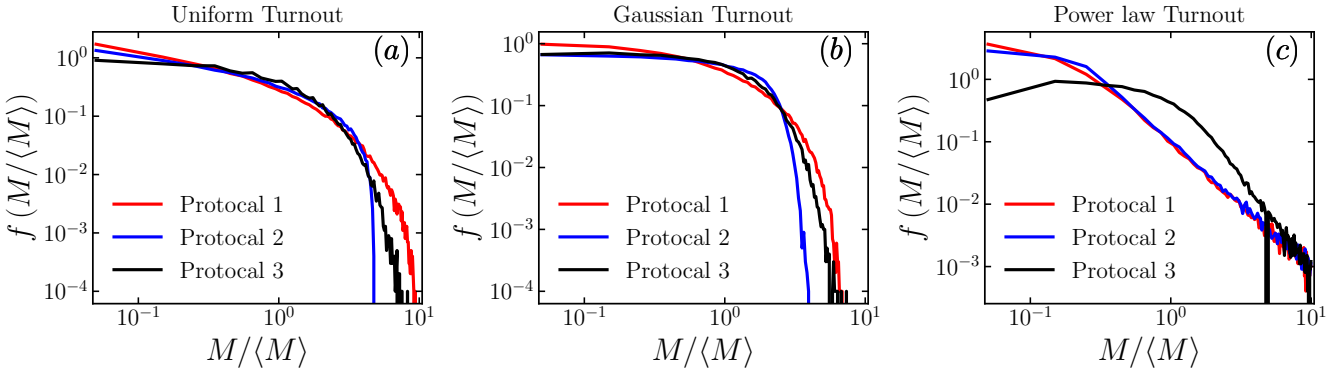

FIG. S2. Prediction of scaled margin distribution for three different protocols of choosing p_i, the probability of receiving votes. Panels (a), (b), and (c) are for uniform, Gaussian, and Power law turnout distributions, respectively.

S6. DATA COLLECTION AND CLEANING

In this work, we use empirical election data from 34 countries. Of these, data from 32 countries are used to establish the universality result, and data from two countries illustrate pronounced cases of deviations from universality.

Data collection– We collect constituency-level data of the lower chamber of the Legislative elections for 180 countries and territories across the world from the Constituency-Level Election Archive (CLEA) website [3]. Polling booth level data for India and Canada is collected from the websites of Election Commission [4, 5] of the respective countries, semi-automatically using a combination of Python libraries. We collect county-level data from MIT Election Data + Science Lab [6] for the USA. While constituency-level data is available for many countries, polling booth-level data is available in the public domain only for a few countries.

Data cleaning–While constituency-level data collected from the CLEA website was in tabular format, the polling booth-level data was found in different formats, ranging from tabular to machine-generated and scanned PDFs. We clean the data using a combination of Python libraries. For each country, election data from multiple constituencies across several elections are aggregated. For instance, if data from 100 constituencies are available for the past five elections, we compile turnouts and margins from each election, resulting in consolidated lists of 500 elements each. These aggregated data are then used for simulations, analysis, and to construct empirical distributions. To ensure a reasonable level of confidence in the statistical analysis, we have ignored data from countries with less than 400 data points. By this criteria, we could use the data from 34 out of 180 countries, all of which have more than 400 data points. The threshold of 400 data points allows us to demonstrate universality, along with flagging possible electoral misconduct in Ethiopia and Belarus, while maintaining good statistics.

In this analysis, we discard those rare cases when the turnout is zero, or the number of contesting candidates is less than two. To avoid discrepancies, we consider the sum of valid votes received by all the candidates (in an electoral unit) as the turnout for the election in that unit. Some important summary statistics of the election data for the 34 countries used for analysis in this work are given in Table S1.

Country	Time span	Number of elections	Scale	Mean turnout	Mean margin	Number of electoral units (consolidated)
Australia	1901-2016	37	Constituency	7.37×10^4	1.31×10^4	1740
Bangladesh	1973-2008	4	Constituency	1.57×10^5	3.15×10^4	1188
Belarus	2004-2019	5	Constituency	4.83×10^4	2.61×10^4	441
Canada	1867-2019	43	Constituency	2.76×10^4	5.50×10^3	10662
Canada	2004-2021	7	Polling Booth	5.56×10^2	1.35×10^2	489919
Chile	1945-2017	7	Constituency	1.07×10^5	1.05×10^4	420
Denmark	1849-2019	30	Constituency	2.70×10^3	4.64×10^2	2178
Ethiopia	2010-2010	1	Constituency	4.95×10^4	4.18×10^4	492
France	1973-2017	3	Constituency	7.88×10^4	1.10×10^4	1712
Germany	1871-2017	19	Constituency	1.37×10^5	2.26×10^4	5108
Ghana	1992-2016	6	Constituency	3.75×10^4	9.88×10^3	1410
Hungary	1990-2018	6	Constituency	5.32×10^4	8.57×10^3	936
India	1951-2019	18	Constituency	5.69×10^5	8.33×10^4	8389
India	2004-2019	4	Polling Booth	5.82×10^2	1.89×10^2	752786
Japan	1947-2017	26	Constituency	2.88×10^5	2.35×10^4	4603
Kenya	1961-2013	2	Constituency	3.72×10^4	1.19×10^4	417
Korea	1948-2012	13	Constituency	6.17×10^4	1.01×10^4	2258
Lithuania	1992-2020	8	Constituency	3.24×10^4	3.98×10^3	570
Malawi	1994-2019	4	Constituency	2.31×10^4	6.29×10^3	755
Malaysia	1959-2018	13	Constituency	3.41×10^4	8.90×10^3	2199
Myanmar	2010-2015	2	Constituency	6.76×10^4	2.32×10^4	634
New Zealand	1943-2020	9	Constituency	3.04×10^4	6.94×10^3	637
Nigeria	2003-2019	2	Constituency	7.75×10^4	2.20×10^4	710
Pakistan	1988-2013	3	Constituency	1.28×10^5	2.45×10^4	683
Papua New Guinea	1972-2017	8	Constituency	5.07×10^4	5.66×10^3	841
Philippines	1946-2013	17	Constituency	1.83×10^5	2.63×10^4	2525
Solomon Islands	1967-2019	14	Constituency	3.67×10^3	4.37×10^2	543
Taiwan	1986-2020	11	Constituency	2.33×10^5	1.98×10^4	482
Tanzania	2005-2020	2	Constituency	5.37×10^4	2.01×10^4	492
Thailand	1969-2011	12	Constituency	1.86×10^5	1.46×10^4	2263
Trinidad and Tobago	1925-2020	13	Constituency	1.53×10^4	5.12×10^3	411
Uganda	2006-2021	4	Constituency	4.45×10^4	1.08×10^4	1430
UK	1832-2019	46	Constituency	3.43×10^4	6.30×10^3	23105
Ukraine	1998-2019	5	Constituency	8.89×10^4	1.67×10^4	1072
United States	1788-2020	167	Congressional District	1.14×10^5	2.96×10^4	33946
United States	2000-2020	6	County	1.78×10^5	2.00×10^4	18905
Zimbabwe	2005-2018	4	Constituency	1.77×10^4	6.55×10^3	743

TABLE S1. Typical values of Margin and turnouts at different scales for different countries. The available data for the mentioned time spans were consolidated for each country and used to calculate the mean turnout and mean margin. The consolidated number of electoral units (in the last column) is calculated by adding the number of valid electoral units for all the elections that happened in the mentioned time span. The data for an electoral unit is considered to be valid if (a) a list of votes received by all the candidates is available, (b) at least two candidates are contesting, and (c) the turnout is non-zero. For example, in polling booth level data for India, lists of votes for all the polling booths are not always available. We could obtain valid data for 752786 polling booths from the four elections held during the time span of $2004 - 2019$ for which the above-mentioned conditions were met. Only national-level elections are considered in this dataset.

S7. FIGURES CONTAINING $f\,(M/\langle M\rangle)$ FOR 32 COUNTRIES

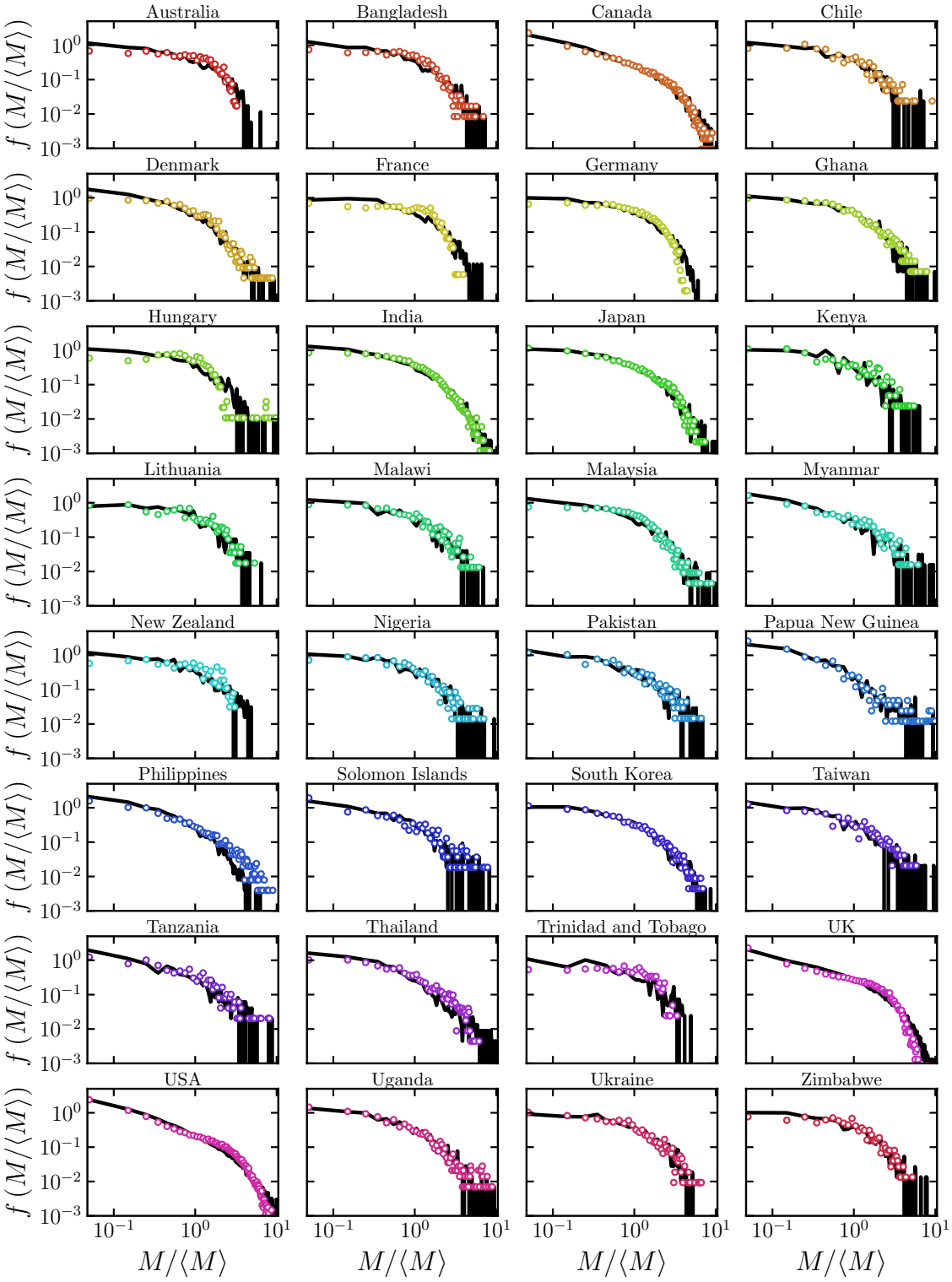

FIG. S3. The empirical distribution of the scaled margins (colored open circles), along with RVM model prediction (black solid lines) for 32 countries.

S8. FIGURES CONTAINING $F(x)$ FOR 32 COUNTRIES

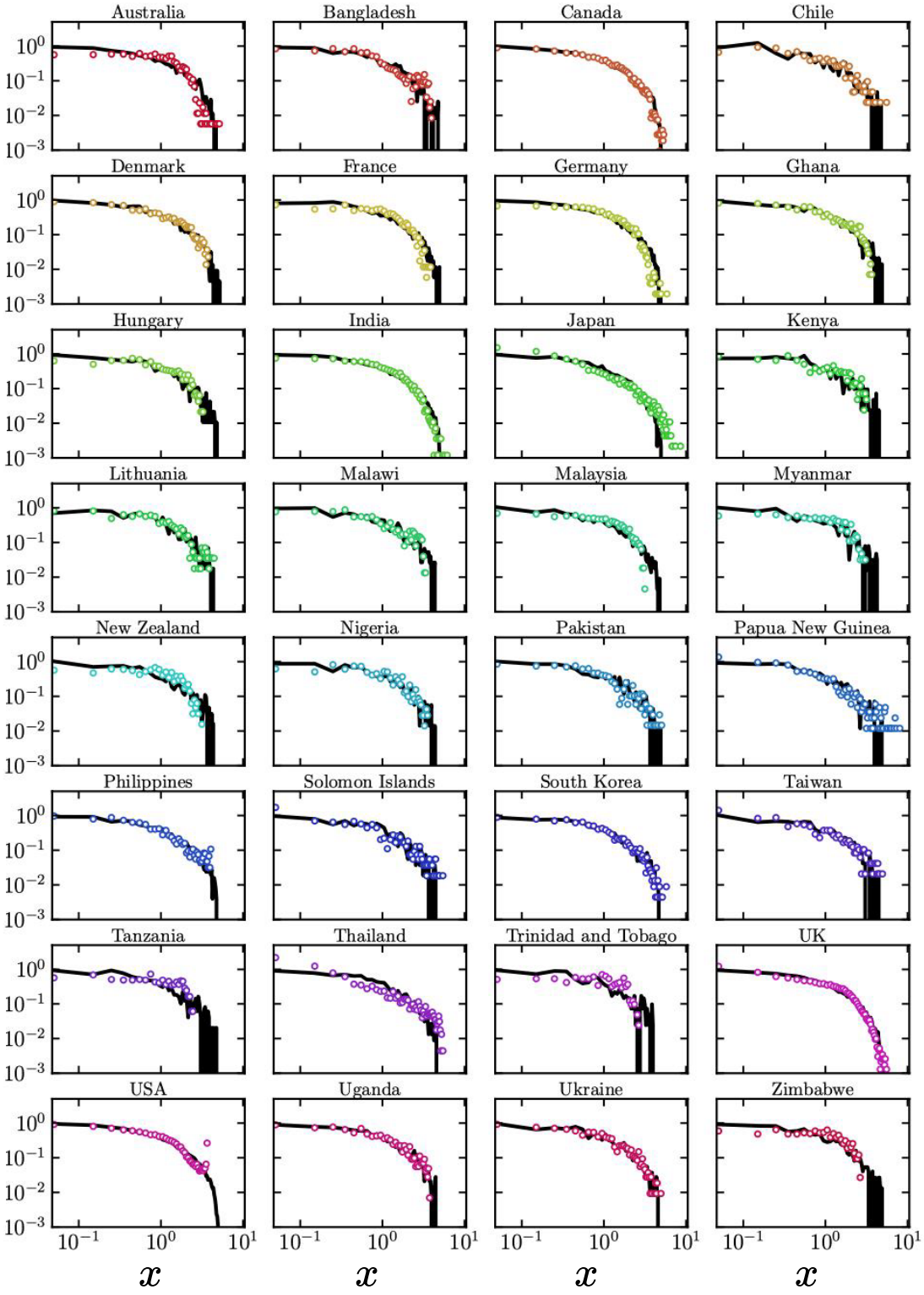

FIG. S4. The empirical distribution of the scaled specific margin (colored open circle), along with RVM model prediction (black solid line) for 32 countries.

S9. SCALING OF $\langle M \rangle$ AND $\langle \mu \rangle$ VS T

At a large turnout limit ($T \gg 1$), the distributions of μ and M produced by RVM are the following,

$$P(\mu) = \frac{(1-\mu)(5+7\mu)}{(1+\mu)^2(1+2\mu)^2},$$
$$\mathcal{P}(M|T) = \frac{(1-M/T)(5+7M/T)}{T(1+M/T)^2(1+2M/T)^2}.$$

From this, we find $\langle \mu \rangle = \frac{1}{2} + \ln\left(\frac{9\sqrt[4]{3}}{16}\right)$ and $\langle M \rangle = T\left(\frac{1}{2} + \ln\left(\frac{9\sqrt[4]{3}}{16}\right)\right)$. We investigate if such linear scaling of $\langle M \rangle$ with T exists in empirical data. As shown in Fig. S5, for some countries (India, Canada, the United States, and the UK), there is a region of linearity in the $\langle M \rangle$ vs T plots. Correspondingly, $\langle \mu \rangle$ is constant in those regions. Japan and Germany pose as counterexamples to this linearity hypothesis, and developing a better understanding of this scaling provides a rich avenue for further research.

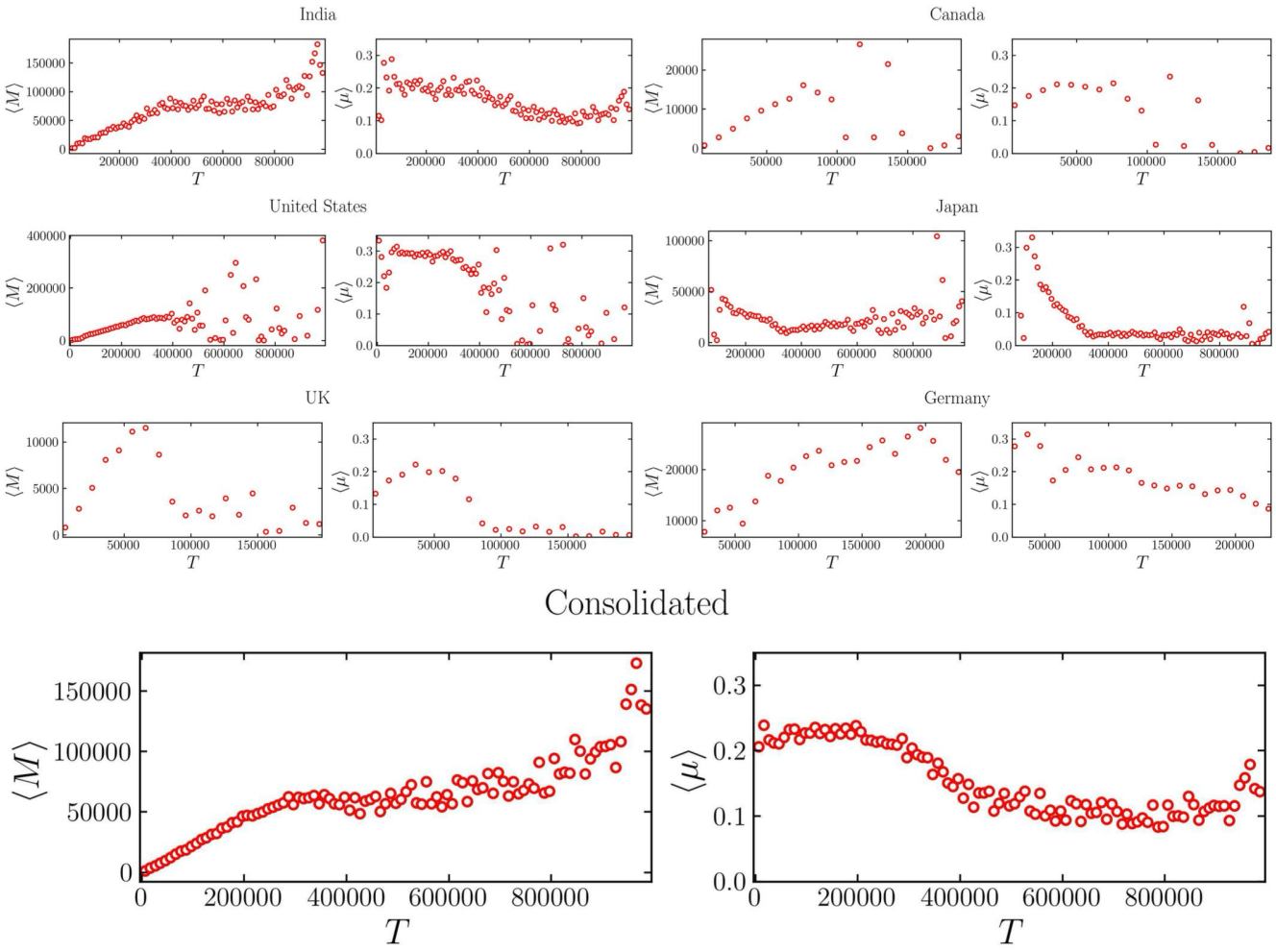

FIG. S5. $\langle M \rangle$ vs T and $\langle \mu \rangle$ vs T for India, Canada, the United States, Japan, the UK, and Germany. The bottom-most panels display plots of the consolidated election data, combining data from 32 countries.

[1] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A First Course in Order Statistics* (Society for Industrial and Applied Mathematics, 2008) https://epubs.siam.org/doi/pdf/10.1137/1.9780898719062.
[2] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, ninth dover printing, tenth gpo printing ed. (Dover, New York, 1964).
[3] "Constituency-level elections archive [data file and codebook]," http://www.electiondataarchive.org.
[4] "Election data of india," https://www.eci.gov.in.
[5] "Election data of canada," https://www.elections.ca.
[6] M. E. Data and S. Lab, "County Presidential Election Returns 2000-2020," (2018).

Voter Turnouts Govern Key Electoral Statistics

Ritam Pal,[1,*] Aanjaneya Kumar,[1,2,3,†] and M. S. Santhanam[1,‡]

[1] *Department of Physics, Indian Institute of Science Education and Research, Pune 411008, India.*
[2] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*
[3] *High Meadows Environmental Institute, Princeton University, Princeton, NJ, 08544, USA*

Elections, the cornerstone of democratic societies, are usually regarded as unpredictable due to the complex interactions that shape them at different levels. In this work, we show that voter turnouts contain crucial information that can be leveraged to predict several key electoral statistics with remarkable accuracy. Using the recently proposed random voting model, we analytically derive the scaled distributions of votes secured by winners, runner-ups, and margins of victory, and demonstrating their strong correlation with turnout distributions. By analyzing Indian election data – spanning multiple decades and electoral scales – we validate these predictions empirically across all scales, from large parliamentary constituencies to polling booths. Further, we uncover a surprising scale-invariant behavior in the distributions of scaled margins of victory, a characteristic signature of Indian elections. Finally, we demonstrate a robust universality in the distribution of the scaled margin-to-turnout ratios.

The institution of election plays a pivotal role in every functioning democracy to ensure that the governing bodies are based on the people's mandate. While the rules for choosing candidates are often simple at the microscopic level, the effects arising from complex interactions among individuals can make electoral processes and their final outcomes unpredictable on larger scales. To address these complexities, tools from statistical physics and complex systems were applied to analyze and uncover patterns in electoral outcomes [1–12].

Over the last few decades, aided by the availability of extensive election data, many earlier studies [13–18] have attempted to identify universal patterns to characterize and simplify the complexities of electoral processes irrespective of microscopic details. While the distributions of vote shares garnered by candidates [19–22] and voter turnouts[23, 24] have been extensively studied, they exhibit limited universality at best [13, 14, 16]. Nevertheless, these distributions have proven valuable in flagging irregularities and detecting fraudulent practices in elections [25–27].

Among the many statistics of interest, the *margin* of victory, defined as $M = V_w - V_r$, where V_w and V_r are the votes secured by the winner and runner-up, respectively, encodes key information about the competitiveness of elections. While margins of victory have been previously studied [28–33], often independently of voter turnouts, our recent work [34] suggests that voter turnouts, in combination with margins, provide deeper insight into electoral dynamics. This extensive analysis of election data spread over many decades of elections held in 34 countries demonstrated that the scaled distribution of the margin-to-turnout ratio exhibits a universal form [34]. Significant deviations from this universal behavior can indicate potential electoral malpractice [34–39]. Furthermore, we demonstrated that voter turnouts play a fundamental role in driving the margins and, together with a proposed random voting model (RVM), can accurately predict the scaled distribution of victory margins. This robust connection between turnouts and margins holds across multiple elections and electoral scales. These findings naturally raise the question: Can the distribution of other relevant electoral statistics be uncovered using voter turnout distributions?

To explore this question, large datasets covering a range of different electoral scales are required. India, the world's largest democracy, regularly conducts elections involving vast electorates (960 million in 2024), and its publicly available election data spans multiple decades and electoral scales. The diversity of India's linguistic and cultural landscape further adds to the complexities of its electoral outcomes, making it an excellent testing ground for assessing the robustness of the RVM framework.

In this Letter, using empirical data from Indian elections [40, 41] spanning several decades and vastly different electoral scales, we demonstrate a strong correlation between the distributions of votes received by winners and runner-ups and voter turnouts. Leveraging this correlation and the RVM, we analytically predict the scaled distributions of the votes secured by winners and runner-ups using the corresponding turnout distributions. This prediction remarkably holds good at all the electoral scales – from large parliamentary constituencies ($\sim 10^6$ voters) down to the smallest polling booth levels ($\sim 10^2 - 10^3$ voters). Further, we show a rather surprising scale invariance of the margin distributions, a characteristic typical of Indian elections. Finally, a robust universality in the distribution of scaled margin-to-turnout ratio is demonstrated, strengthening our recent proposition.

We formalize our framework as follows: An *election* happens at all the N electoral units following the first-post-the-past principle [42]. Let the i-th electoral

* ritam.pal@students.iiserpune.ac.in
† aanjaneya@santafe.edu
‡ santh@iiserpune.ac.in

unit have n_i^c candidates and n_i^v eligible voters, where $i = 1, 2, \ldots N$. Usually, only a fraction of the eligible voters cast their votes. This is termed the turnout $T_i \leq n_i^v$. It is a direct indicator of the people's interest in the electoral process, and its distribution $g(T)$ encodes information about electoral statistics [34]. Let $v_{i,1}, v_{i,2}, \ldots, v_{i,n_i^c}$ be the votes secured by n_i^c candidates such that $\sum_j v_{i,j} = T_i$. The candidate securing the highest number of votes, V_w, is declared the winner, while the candidate with the second-highest votes, V_r, is the runner-up. By definition $V_w > V_r$. Further, the *margin* of victory is defined as $M_i = V_{i,w} - V_{i,r}$, and indicates the extent of electoral competition.

The electoral units in this work have three distinct scales in terms of the size of the electorate: (i) parliamentary constituency (PC, largest scale), (ii) assembly constituency (AC, intermediate scale), and (iii) polling booth (PB, smallest scale). Table I describes the granularity of election data used in this work – electoral units and their typical electorate size. While for the national-level general elections, we employ data from 3 different electoral scales (PC, AC, and PB), we use AC-level data for state elections. Depending on the electoral level considered, the winner and runner-up vote distributions have vastly different scales, with the winner vote distribution having wider support than the runner-up. However, when both distributions are scaled by their respective mean values (mean taken over all elections for which data is available), the winner and runner-up vote distributions – $Q_{\widetilde{V_w}}(\widetilde{V_w})$ and $Q_{\widetilde{V_r}}(\widetilde{V_r})$ – explicitly display a strong correlation with the corresponding scaled turnout distributions $Q_{\widetilde{T}}(\widetilde{T})$. Note that any variable Y, scaled by their mean $\langle Y \rangle$, is denoted by $\widetilde{Y} = Y/\langle Y \rangle$. Figure 1 displays this key result at four electoral units described in Table I. Remarkably, at larger electoral scales (PC and AC), not only the tail but the entire scaled distributions of winner and runner-up votes mimic the corresponding scaled turnout distribution as seen in Fig. 1(b-d). This strong correlation indicates that the turnout distribution contains crucial information about different election statistics and can be leveraged to predict the scaled vote distributions of the winner and the runner-up. To explore this possibility, we employ our recently proposed random voting model (RVM) [34], which is demonstrably effective at predicting the scaled distribution of several election statistics, such as the *margin* of victory.

The random voting model, later denoted as RVM (T, n^c), is built around the framework of elections described above and consists of N electoral units with n_i^c number of candidates contesting for the votes of T_i electors who have voted in the i-th electoral unit. Then, the probability that j-th candidate attracts electors' votes is:

$$p_{ij} = \frac{w_{ij}}{\sum_{k=1}^{n_i^c} w_{ik}}, \quad w_{ij} \sim \mathcal{U}(0,1), \quad (j = 1, 2, \cdots n_i^c). \tag{1}$$

In this, $\mathcal{U}(0,1)$ is the uniform distribution. This model was shown to capture the various statistical features of empirical election data, such as the margin distributions and the universal features embedded in the election data [34]. In particular, the random voting model with $n^c = 3$ candidates – RVM $(T, 3)$ – predicts the scaled distribution of *margin* remarkably well, irrespective of electoral scales and countries [34]. In this work, we employ a refined approach based on the notion of the effective number of parties defined as [43]:

$${}^{(E)}n_i^c = \frac{1}{\sum_{k=1}^{n_i^c} (V_{ik}/T_i)^2}. \tag{2}$$

In large elections exercise such as that in India, even though many candidates join the fray, a few corner most of the votes. For instance, if all the votes are garnered by just one candidate, then $V_{i1} = T_i$, and $V_{ij} = 0$ for $j = 2, \ldots n^c$. In this case, ${}^{(E)}n_i^c = 1$. However, if all the votes are split equally among two candidates, ${}^{(E)}n_i^c = 2$, thus, Eq. 2 captures the idea of an effective number of candidates in i-th electoral unit. Further, by averaging over all the electoral units, we obtain:

$${}^{(E)}\tilde{n}^c = \left[\frac{1}{N} \sum_{k=1}^{N} {}^{(E)}n_k^c \right], \tag{3}$$

where $[*]$ denotes the operation of extracting the closest integer value. And ${}^{(E)}\tilde{n}^c$ indicates the effective number

TABLE I. Electoral units in various elections in India

Type	Type of electoral unit	Type of election	Size
GE-PB	Polling booth	General election	10^3
GE-AC	Assembly constituency	General election	10^5
GE-PC	Parliamentary constituency	General election	10^6
SE-AC	Assembly constituency	State election	10^5

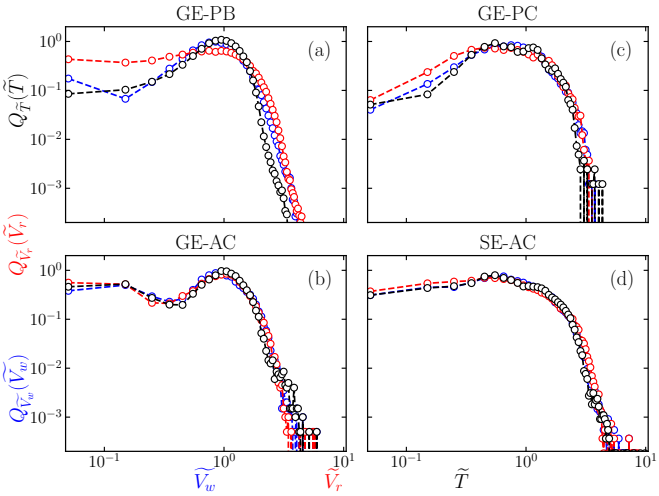

FIG. 1. Winner, runner-up vote distributions, and turnout distributions, scaled by their respective means. Notably, at larger electoral scales (AC / PC), the winner and runner-up distributions mimic the corresponding turnout distribution.

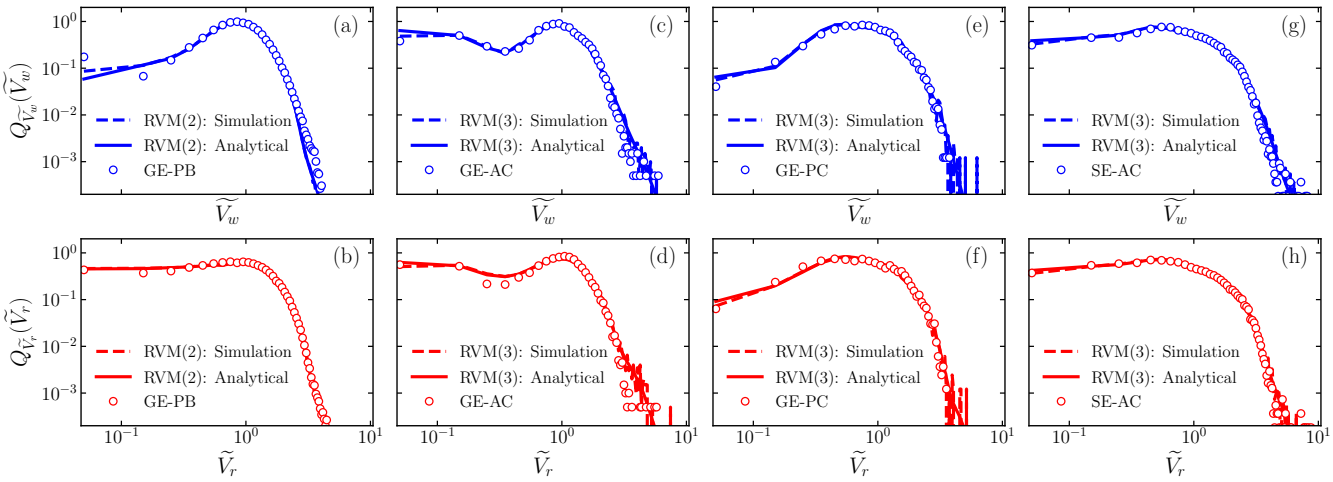

FIG. 2. Winner and runner-up vote distributions scaled by their respective means. Panels (a, b), (c, d), and (e, f) depict, respectively, the scaled winner and runner-up vote distribution at the polling booth, assembly constituency, and parliamentary constituency level for Indian general elections. Panels (g, h) correspond to the distributions for the state elections at the assembly constituency level. The analytical predictions (solid lines) are in remarkable agreement with the empirical distributions (open circle). Predictions from RVM simulations (dashed line) closely follow the analytical curves.

of candidates for an entire election at different electoral scales. From empirical data of Indian elections at different scales, we find ${}^{(E)}\tilde{n}^c = 2$ at the polling booth (PB-GE) level for the General Elections. However, for all the other three cases of AC-GE, PC-GE, and AC-SE, we obtain ${}^{(E)}\tilde{n}^c = 3$. Now, we shall solve RVM (T, n^c) for $n^c = 2$ and $n = 3$ to analytically describe the results observed in Fig. 1.

The primary object of interest is the distribution of the votes received by the winner and the runner-up. In the large turnout limit, $T >> 1$, the votes received by j-th candidate can be approximated as $V_j \approx p_j T$ (index i is dropped as we focus on a single electoral unit). Consequently, the vote share is defined as:

$$v_j = V_j/T \quad \text{with } j = 1, 2, \dots n^c. \tag{4}$$

Thus, in this limit, the vote share distribution is effectively the same as the distribution of p_j. Recall that p_j can be constructed from random weights using Eq. 1. These weights are n^c *i.i.d.* random variables $\{w_1, w_2, \cdots, w_{n^c}\}$ drawn from the uniform distribution $\mathcal{U}(0, 1)$. When arranged in ascending order, the random variable at the k-th place is defined as the k-th order statistics and is denoted by $w^{n^c}_{(k)}$ [44]. Specifically, $w^{n^c}_{(1)}$ and $w^{n^c}_{(n^c)}$ represent the smallest and the largest weights, respectively. Then, the joint probability distribution function (jpdf) or all the order statistics is given by:

$$\mathbb{P}\left(w^{n^c}_{(1)}, w^{n^c}_{(2)}, \dots w^{n^c}_{(n^c)}\right) = n^c!. \tag{5}$$

Finally, the winner's vote share v_w and runner-up's vote share v_r can be expressed in terms of order statistics as:

$$v_w = \frac{w^{n^c}_{(n^c)}}{\sum_{k=1}^{n^c} w^{n^c}_{(k)}} \quad \text{and} \quad v_r = \frac{w^{n^c}_{(n^c-1)}}{\sum_{k=1}^{n^c} w^{n^c}_{(k)}}. \tag{6}$$

Their distributions can be obtained from the jpdf in Eq. 5 by integrating out the other variables (for detailed calculations, see Supplementary Material [45]). When the number of candidates $n^c = 2$, the distribution of the winner's vote share $P_{v_w}(v_w)$ is found to be:

$$P_{v_w}(v_w) = \begin{cases} \dfrac{1}{v_w^2}, & \text{if} \quad \dfrac{1}{2} < v_w < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The vote share distribution of the runner-up can also be calculated similarly and is given by:

$$P_{v_r}(v_r) = \begin{cases} \dfrac{1}{(1-v_r)^2}, & \text{if} \quad 0 < v_r \leq \dfrac{1}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The conditions in Eqs. 7-8 reflect the intuitive idea that when there are only two candidates, the winner's vote share cannot be less than 1/2, and the runner-up cannot exceed 1/2. For $n^c = 3$, the winner's vote share distribution is:

$$P_{v_w}(v_w) = \begin{cases} \dfrac{3v_w - 1}{v_w^3}, & \text{if} \quad \dfrac{1}{3} < v_w \leq \dfrac{1}{2} \\ \dfrac{1 - v_w}{v_w^3}, & \text{if} \quad \dfrac{1}{2} < v_w < 1 \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

and the distribution of the runner-up's vote share is:

$$P_{v_r}(v_r) = \begin{cases} \dfrac{v_r(2-3v_r)}{(1-v_r)^2(1-2v_r)^2}, & \text{if} \quad 0 < v_r \leq \dfrac{1}{3} \\ \dfrac{1-2v_r}{v_r^2(1-v_r)^2}, & \text{if} \quad \dfrac{1}{3} < v_r \leq \dfrac{1}{2} \\ 0, \quad \text{otherwise.} \end{cases} \tag{10}$$

The vote share distribution of the winner and the runner-up are defined as piecewise functions. They are non-zero when $\frac{1}{3} < v_w < 1$ and $0 < v_r \leq \frac{1}{2}$, respectively. Based on Eq. 4, the distribution of unscaled variables $Y = (V_w, V_r)$, given T, is related to the scaled variables $y = (v_w, v_r)$ via:

$$\mathcal{P}(Y|T) = \frac{1}{T} P_y\left(\frac{Y}{T}\right), \tag{11}$$

where $P_y(y)$ is the probability density function for the scaled variable, $y = (v_w, v_r)$. The distribution of Y for arbitrary turnout distribution $g(T)$ is:

$$Q_Y(Y) = \int g(T)\ \mathcal{P}(Y|T)\ dT, \tag{12}$$

with $\langle Y \rangle = \int Y\ Q_Y(Y)\ dY$. Finally, for the scaled variable $\widetilde{Y} = Y/\langle Y \rangle$, the distribution is:

$$Q_{\widetilde{Y}}(\widetilde{Y}) = \langle Y \rangle\ Q_Y(\widetilde{Y}\langle Y \rangle). \tag{13}$$

Using the empirical turnout distribution $g(T)$ from election data, Eq. 12 is numerically integrated. The resulting distribution is then scaled using Eq. 13 to obtain the scaled distributions for the winner and runner-up vote shares, $Q_{\widetilde{V_w}}(\widetilde{V_w})$ and $Q_{\widetilde{V_r}}(\widetilde{V_r})$, respectively. As demonstrated in Fig. 2, the analytical prediction (solid lines) is remarkably consistent with the empirical vote share distributions. The predictions from RVM simulations, which use the raw turnout data and $n^c = {}^{(E)}\tilde{n}^c$ as inputs, closely follow the analytical distributions in Fig. 2. The scaled distributions of winner and runner-up votes depicted across all electoral scales, in Fig. 2, typically exhibit a power-law behavior in the tails for $\widetilde{V}_w, \widetilde{V}_r \gg 1$. Conversely, for $\widetilde{V}_w, \widetilde{V}_r \ll 1$, the distributions display different profiles. Remarkably, these differences are well captured by RVM predictions: RVM $(T, 2)$ accurately predicts distribution at the GE-PB level, while RVM $(T, 3)$ closely matches the distributions at the GE-AC, GE-PC, and SE-AC levels. Hence, the effective number of candidates (Eq. 2) and the turnout distribution $g(T)$, when used within the RVM framework, successfully predict the winner and runner-up vote share distributions across distinct electoral scales.

Next, we consider the effect of voter turnouts T on the *margin* of victory M. To do this, firstly we define *specific margin* as $\mu = M/T = (V_w - V_r)/T$. In the large turnout $(T >> 1)$ limit, the specific margin can be expressed in terms of the order statistics of w as:

$$\mu = \frac{w^{n^c}_{(n^c)} - w^{n^c}_{(n^c-1)}}{\sum_{k=1}^{n^c} w^{n^c}_{(k)}}, \tag{14}$$

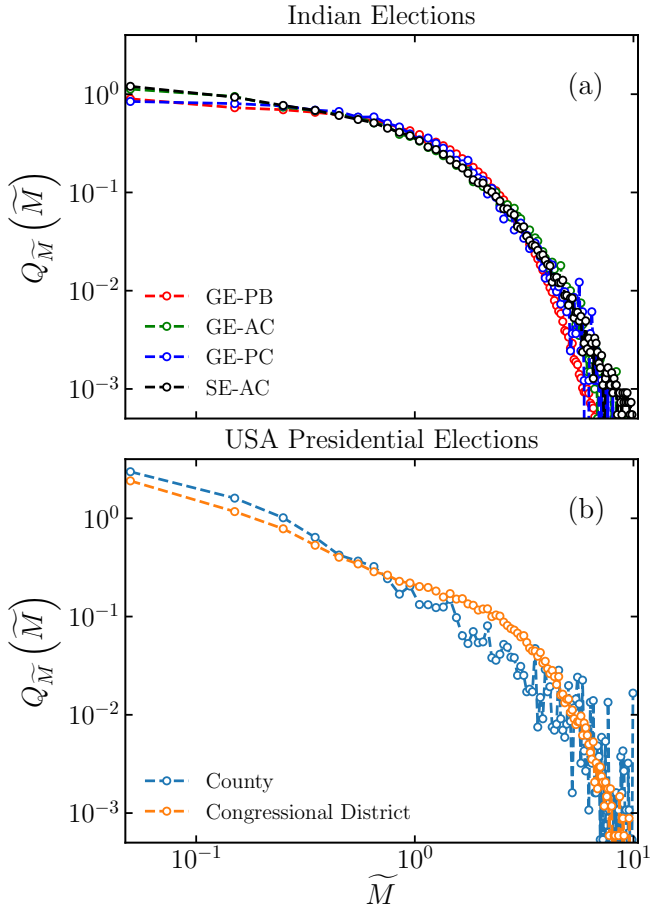

FIG. 3. Margin distributions scaled by their respective means. (a) Data collapse in the scaled margin distributions of Indian elections at four electoral scales. (b) In contrast, such collapse is absent in the election data from the USA.

where n^c is the number of candidates. Using the jpdf in Eq. 5, for RVM $(T, 2)$ the distribution of specific margin can be obtained as (see Ref. [45] for more details):

$$P_\mu(\mu) = \frac{2}{(1+\mu)^2}, \tag{15}$$

and for RVM $(T, 3)$, the distribution becomes:

$$P_\mu(\mu) = \frac{(1-\mu)(5+7\mu)}{(1+\mu)^2(1+2\mu)^2}. \tag{16}$$

Using Eq. 12 - 13 and the empirical turnout distributions $g(T)$, the scaled margin distribution $Q_{\widetilde{M}}(\widetilde{M})$ can be obtained for the four different electoral scales which match the corresponding empirical distributions closely (see Ref. [45]). Remarkably, the scaled distributions for the margin for Indian elections at four different scales collapse onto a single curve, as shown in Fig. 3(a). This data collapse is a direct consequence of the similarity in tail behavior in the corresponding turnout distributions (see Fig. 1). This appears to be a characteristic of Indian elections and is not observed in most other countries. For

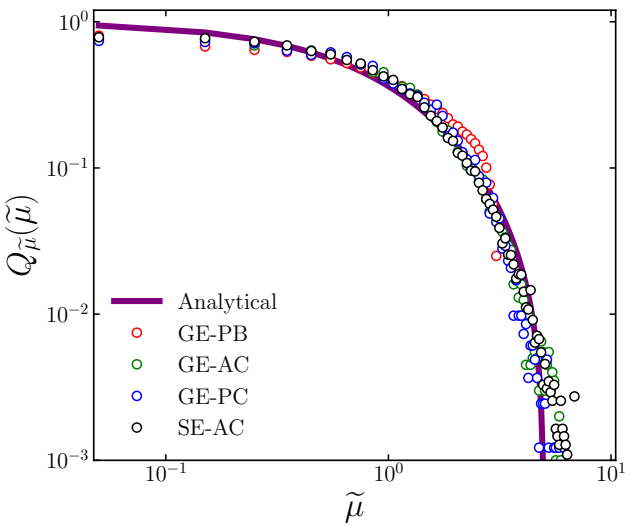

FIG. 4. Specific margin distributions scaled by their respective means at distinct electoral scales in different types of Indian elections. The scaled specific margin distributions collapse on the analytical universal curve.

instance, this data collapse is absent in the US elections for the empirical data at the County and Congressional district levels, as demonstrated in Fig. 3 (b).

Recently, we have shown that the scaled distribution of specific margins $Q_{\widetilde{\mu}}(\widetilde{\mu})$, is universal irrespective of electoral scales and countries [34]. This universal distribution can be analytically obtained by rescaling Eq. 16 with $\langle\mu\rangle = 1/2 + \ln\left(9\sqrt[4]{3}/16\right)$ (see Ref. [45]). The empirical distributions from large election datasets are proposed to exhibit a better collapse to the analytical curve, as finite-size effects are suppressed. The presence of such universality helps us distill the complexity of the electoral processes in terms of universal behavior. As the electorate size in India is large (960 million in 2024) and datasets are available at various electoral scales, Indian elections provide the best test-bed to demonstrate such universality. Fig. 4 demonstrates such universality. The distributions $Q_{\widetilde{\mu}}(\widetilde{\mu})$ at four distinct electoral scales (colored open circles) remarkably collapse onto the analytically predicted universal curve (solid line). This excellent collapse strengthens the proposition that in fairly conducted elections, the apparent deviation from universality can originate from finite data size.

In summary, elections are a source of excellent datasets for exploring collective decision-making by millions of people (interacting agents) at *distinct electoral scales*. However, most of the earlier works on elections have ignored the effects arising from differences in electoral scales. The voter turnout distribution $g(T)$ is a good indicator of the public trust and interest in the electoral process. Remarkably, we show that they also encode information about several crucial election statistics. Aided by election data from India – the largest democracy in the world – we demonstrate our results for different types of elections at multiple *distinct electoral scales*. Given the empirical $g(T)$ and an effective number of candidates for each electoral scale, we analytically obtain the scaled distributions of the winner and the runner-up votes using the framework of our recently proposed random voting model. The analytical predictions and the simulations are in excellent agreement with the empirical election data. Further, we demonstrate that the random voting model is effective in predicting the scaled margin distributions of Indian elections at vastly different electoral scales. Surprisingly, the scaled margin distribution remains invariant with respect to changes in electoral scales, making it a characteristic feature of Indian elections. Finally, the scaled specific margin distributions of Indian elections show a remarkable data collapse, strengthening the recently proposed universality [34]. This work paves the way for a wider understanding of electoral statistics from turnout distributions.

ACKNOWLEDGMENTS

R.P. and A.K. thank the Prime Minister's Research Fellowship of the Government of India for financial support. M.S.S. acknowledges the support of a MATRICS Grant from SERB, Government of India, during the early stages of this work. The authors acknowledge the National Supercomputing Mission for the use of PARAM Brahma at IISER Pune.

[1] S. Galam, Application of statistical physics to politics, Physica A: Statistical mechanics and its applications **274**, 132 (1999).
[2] A. Gelman, J. N. Katz, and F. Tuerlinckx, The mathematics and statistics of voting power, Statistical Science , 420 (2002).
[3] S. J. Brams, *Mathematics and Democracy : Designing better voting and fair-division procedures* (Princeton University Press, Princeton, 2008).
[4] C. Castellano, S. Fortunato, and V. Loreto, Statistical physics of social dynamics, Rev. Mod. Phys. **81**, 591 (2009).
[5] S. Galam, *Sociophysics: A Physicist's Modeling of Psycho-political Phenomena* (Springer New York, NY, 2012).
[6] S. Fortunato, Santo fortunato· michael macy· sidney redner, J Stat Phys **151**, 1 (2013).
[7] P. Sen and B. K. Chakrabarti, *Sociophysics: an introduction* (OUP, Oxford, 2014).
[8] J. Fernández-Gracia, K. Suchecki, J. J. Ramasco, M. San Miguel, and V. M. Eguíluz, Is the voter model a model for voters?, Phys. Rev. Lett. **112**, 158701 (2014).

[9] D. Braha and M. A. M. de Aguiar, Voting contagion: Modeling and analysis of a century of u.s. presidential elections, PLOS ONE **12**, 1 (2017).

[10] A. Kononovicius, Compartmental voter model, Journal of Statistical Mechanics: Theory and Experiment **2019**, 103402 (2019).

[11] S. Redner, Reality-inspired voter models: A mini-review, Comptes Rendus Physique **20**, 275 (2019).

[12] M. San Miguel and R. Toral, Introduction to the chaos focus issue on the dynamics of social systems, Chaos: An Interdisciplinary Journal of Nonlinear Science **30** (2020), see all the papers that are part of this special issue.

[13] R. N. C. Filho, M. P. Almeida, J. S. Andrade, and J. E. Moreira, Scaling behavior in a proportional voting process, Phys. Rev. E **60**, 1067 (1999).

[14] S. Fortunato and C. Castellano, Scaling and universality in proportional elections, Phys. Rev. Lett. **99**, 138701 (2007).

[15] M. Mantovani, H. Ribeiro, M. Moro, S. Picoli, and R. Mendes, Scaling laws and universality in the choice of election candidates, Europhysics Letters **96**, 48001 (2011).

[16] A. Chatterjee, M. Mitrović, and S. Fortunato, Universality in voting behavior: an empirical analysis, Scientific reports **3**, 1049 (2013).

[17] E. Bokányi, Z. Szállási, and G. Vattay, Universal scaling laws in metro area election results, PLOS ONE **13**, 1 (2018).

[18] V. Hösel, J. Müller, and A. Tellier, Universality of neutral models: Decision process in politics, Palgrave Communications **5**, 1 (2019).

[19] A. M. Calvão, N. Crokidakis, and C. Anteneodo, Stylized facts in brazilian vote distributions, PLOS ONE **10**, 1 (2015).

[20] K. Burghardt, W. Rand, and M. Girvan, Competing opinions and stubborness: Connecting models to data, Phys. Rev. E **93**, 032305 (2016).

[21] S. Mori, M. Hisakado, and K. Nakayama, Voter model on networks and the multivariate beta distribution, Phys. Rev. E **99**, 052307 (2019).

[22] A. Kononovicius, Modeling of the parties' vote share distributions, Acta Physica Polonica A **133**, 1450 (2018).

[23] C. Borghesi and J.-P. Bouchaud, Spatial correlations in vote statistics: a diffusive field model for decision-making, Eur. Phys. J. B **75**, 395 (2010).

[24] C. Borghesi, J.-C. Raynal, and J.-P. Bouchaud, Election turnout statistics in many countries: Similarities, differences, and a diffusive field model for decision-making, PLOS ONE **7**, 1 (2012).

[25] P. Klimek, Y. Yegorov, R. Hanel, and S. Thurner, Statistical detection of systematic election irregularities, Proceedings of the National Academy of Sciences **109**, 16469 (2012).

[26] R. Jimenez, M. Hidalgo, and P. Klimek, Testing for voter rigging in small polling stations, Science advances **3**, e1602363 (2017).

[27] A. Rozenas, Detecting election fraud from irregularities in vote-share distributions, Political Analysis **25**, 41 (2017).

[28] G. C. Jacobson, The marginals never vanished: Incumbency and competition in elections to the us house of representatives, 1952-82, American Journal of Political Science , 126 (1987).

[29] S. J. McCann, Threatening times," strong" presidential popular vote winners, and the victory margin, 1824–1964., Journal of personality and social psychology **73**, 160 (1997).

[30] C. B. Mulligan and C. G. Hunter, The empirical frequency of a pivotal vote, Public Choice **116**, 31 (2003).

[31] T. R. Magrino, R. L. Rivest, and E. Shen, Computing the margin of victory in {IRV} elections, in *2011 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE 11)* (2011).

[32] L. Xia, Computing the margin of victory for various voting rules, in *Proceedings of the 13th ACM conference on electronic commerce* (2012) pp. 982–999.

[33] A. Bhattacharyya and P. Dey, Predicting winner and estimating margin of victory in elections using sampling, Artificial Intelligence **296**, 103476 (2021).

[34] R. Pal, A. Kumar, and M. Santhanam, Universal statistics of competition in democratic elections, arXiv preprint arXiv:2401.05065 (2024).

[35] G. Brigaldino, Elections in the imperial periphery: Ethiopia hijacked, Review of African political economy **38**, 327 (2011).

[36] M. Frear, The parliamentary elections in belarus, september 2012, Electoral Studies **33**, 350 (2014).

[37] Report of organization for security and co-operation in europe (osce), `https://www.osce.org/odihr/elections/belarus` (2020).

[38] A. Czwołek and J. Kołodziejska, Belarusian parliamentary election in 2019, The Copernicus Journal of Political Studies , 81 (2021).

[39] S. Bedford, The 2020 presidential election in belarus: Erosion of authoritarian stability and re-politicization of society, Nationalities Papers **49**, 808–819 (2021).

[40] Election data of india, `https://www.eci.gov.in`.

[41] Lok dhaba - a repository of indian election results, `https://lokdhaba.ashoka.edu.in/`.

[42] L. Johnston, *Politics: An introduction to the modern democratic state* (University of Toronto Press, 2007).

[43] M. Laakso and R. Taagepera, "effective" number of parties: a measure with application to west europe, Comparative political studies **12**, 3 (1979).

[44] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A First Course in Order Statistics* (Society for Industrial and Applied Mathematics, 2008).

[45] See Supplemental Material [URL] for (1) the description of RVM, (2) theoretical calculations for RVM and other related discussions, (3) data summary, and (4) figures.

Supplemental Material for "Voter Turnouts Govern Key Electoral Statistics"

This Supplemental Material provides further discussion and derivations which support the findings reported in the Letter, and provides details of the models and simulations used to validate the results.

CONTENTS

S1. DESCRIPTION OF RANDOM VOTING MODEL (RVM)

In the Random Voting Model, N electoral units are considered, with n_i^c number of candidates contesting to win votes from T_i voters present to cast their votes in the i-th electoral unit. Each of the n_i^c candidates is assigned a random weight w_{ij}. These weights are drawn independently from a uniform distribution between 0 and 1. The corresponding probability p_{ij} of receiving votes is calculated by normalizing these weights. Hence, we have the following,

$$w_{ij} \sim \mathcal{U}(0,1) \text{ and } p_{ij} = \frac{w_{ij}}{\sum_{k=1}^{3} w_{ik}};\ \text{with } j = 1, 2 \cdots n_i^c. \tag{S1}$$

For the rest of the analysis, we focus on a single (i-th) electoral unit with voter turnout T and drop the corresponding index i for brevity. Hence,

$$w_{ij} := w_j \text{ and } p_{ij} := p_j. \tag{S2}$$

S2. ANALYTICAL CALCULATIONS OF DIFFERENT ELECTION STATISTICS

For large turnout ($T \gg 1$), it is reasonable to assume the number of votes received by j-th candidate is proportional to their probability p_j, in particular, $V_j \approx p_j T$. Hence, for $T \gg 1$, the votes received by the winner V_w can be approximated as,

$$V_w \approx p_{max} T, \tag{S3}$$

and the votes received by the runner-up V_r as,

$$V_r \approx p_{2nd\,max} T, \tag{S4}$$

and the margin $M = V_w - V_r$ can also be written as,

$$M \approx (p_{max} - p_{2nd\,max})\,T, \tag{S5}$$

where p_{max} and $p_{2nd\,max}$ correspond to the largest and the second largest probabilities assigned to the candidates. For example, if the number of candidates $n^c = 3$ and the probabilities p_1, p_2, and p_3 assigned to those 3 candidates are $0.2, 0.5$, and 0.3, then $p_{max} = p_2 = 0.5$ and $p_{2nd\,max} = p_3 = 0.3$.

A. Order statistics and its connection to winner and runner-up vote share and margins

Consider n *iid* random variables $\{X_1, X_2 \dots X_n\}$ drawn from a distribution $\rho(x)$. When arranged in ascending order, the random variable at the k-th spot is defined as the k-th order statistics. In particular, n-th and 1-st order statistics correspond to the maximum and minimum of those n random variables, respectively. The k-th order statistics of the random variable X is denoted by $X_{(k)}$.

The joint probability density of all the order statistics of the above-mentioned n random variables, $\mathbb{P}\left(x_{(1)}, x_{(2)}, ...x_{(n)}\right)$, defined as the probability density that the random variable $X_{(k)}$ takes the value $x_{(k)}$ for $k \in \{1, 2, \dots, n\}$, is

$$\mathbb{P}\left(x_{(1)}, x_{(2)}, ...x_{(n)}\right) = n! \prod_{k=1}^{n} \rho\left(x_{(k)}\right). \tag{S6}$$

Now as described in Eq. S1 the probabilities p_j can be expressed in terms of w_j. Hence the winner vote share $v_w = V_w/T$, runner-up vote share $v_r = V_r/T$ and the specific margin $\mu = M/T$ can be expressed in terms of w as the following,

$$v_w = \frac{V_w}{T} \approx p_{max} = \frac{w_{max}}{\sum_{k=1}^{n^c} w_k} = \frac{w_{(n^c)}}{\sum_{k=1}^{n^c} w_{(k)}} \tag{S7}$$

$$v_r = \frac{V_r}{T} \approx p_{2nd\,max} = \frac{w_{2nd\,max}}{\sum_{k=1}^{n^c} w_k} = \frac{w_{(n^c-1)}}{\sum_{k=1}^{n^c} w_{(k)}} \tag{S8}$$

$$\mu = \frac{M}{T} \approx p_{max} - p_{2nd\,max} = \frac{w_{max} - w_{2nd\,max}}{\sum_{k=1}^{n^c} w_k} = \frac{w_{(n^c)} - w_{(n^c-1)}}{\sum_{k=1}^{n^c} w_{(k)}}, \tag{S9}$$

where $w_{(k)}$ is the k-th order statistics [44].

B. Random Voting Model with two candidates

In the two-candidate Random Voting Model, we have $n = n^c = 2$ and $p(x) = \mathcal{U}(0, 1)$. Hence, the joint probability distribution of all the order statistics have the following form,

$$\mathbb{P}\left(w_{(1)}, w_{(2)}\right) = 2! = 2; \text{ with } 0 < w_{(1)} < w_{(2)} < 1, \tag{S10}$$

and $\mathbb{P}\left(w_{(1)}, w_{(2)}\right) = 0$ otherwise, with the following normalization:

$$\int_0^1 dw_{(2)} \int_0^{w_{(2)}} 2dw_{(1)} = 1. \tag{S11}$$

1. *Winner vote share distribution*

From the joint probability distribution of all the order statistics (Eq. S10), the approximate vote share distribution of the winner can be obtained as,

$$P_{v_w}(v_w) = 2\int_0^1 dw_{(2)} \int_0^{w_{(2)}} \delta\left(v_w - \frac{w_{(2)}}{w_{(1)}+w_{(2)}}\right) dw_{(1)},$$
$$= 2\int_0^1 \frac{2w_{(2)}}{v_w^2} \mathbb{1}_{1/2\le v_w<1} dw_{(2)}, \tag{S12}$$

or,

$$P_{v_w}(v_w) = \begin{cases} \dfrac{1}{v_w^2} \text{ if } \dfrac{1}{2} \le v_w < 1 & \text{(S13)} \\ 0, \text{ otherwise.} & \text{(S14)} \end{cases}$$

2. *Runner-up vote share distribution*

We can similarly calculate the probability density function of the runner-up vote share as the following,

$$P_{v_r}(v_r) = 2\int_0^1 dw_{(2)} \int_0^{w_{(2)}} \delta\left(v_r - \frac{w_{(1)}}{w_{(1)}+w_{(2)}}\right) dw_{(1)},$$
$$= 2\int_0^1 \frac{2w_{(2)}}{(1-v_r)^2} \mathbb{1}_{0<v_r<1/2} dw_{(2)}, \tag{S15}$$

or,

$$P_{v_r}(v_r) = \begin{cases} \dfrac{1}{(1-v_r)^2} \text{ if } 0 < v_r < \dfrac{1}{2} & \text{(S16)} \\ 0, \text{ otherwise.} & \text{(S17)} \end{cases}$$

3. *Specific margin distribution*

Similarly, the distribution of the specific margin $\mu = M/T$ can be obtained as,

$$P_{\mu}(\mu) = 2\int_0^1 dw_{(2)} \int_0^{w_{(2)}} \delta\left(\mu - \frac{w_{(2)}-w_{(1)}}{w_{(1)}+w_{(2)}}\right) dw_{(1)},$$
$$= 2\int_0^1 \frac{4w_{(2)}}{(1+\mu)^2} dw_{(2)}, \tag{S18}$$

or,

$$P_{\mu}(\mu) = \begin{cases} \dfrac{2}{(1+\mu)^2} \text{ if } 0 < \mu < 1 & \text{(S19)} \\ 0, \text{ otherwise.} & \text{(S20)} \end{cases}$$

C. Random Voting Model with three candidates

In the three-candidate Random Voting Model, we have $n = n^c = 3$ and $p(x) = \mathcal{U}(0,1)$. Then, the joint probability distribution of all the order statistics is,

$$\mathbb{P}\left(w_{(1)}, w_{(2)}, w_{(3)}\right) = 3! = 6; \text{ with } 0 < w_{(1)} < w_{(2)} < w_{(3)} < 1, \tag{S21}$$

and $\mathbb{P}\left(w_{(1)}, w_{(2)}, w_{(3)}\right) = 0$ otherwise, with the following normalization:

$$\int_0^1 dw_{(3)} \int_0^{w_{(3)}} dw_{(2)} \int_0^{w_{(2)}} 6dw_{(1)} = 1. \tag{S22}$$

1. Winner vote share distribution

From the joint probability distribution of all the order statistics, we calculate the approximate probability density function of the winner vote share $v_w = V_w/T$ as follows,

$$\begin{aligned} P_{v_w}(v_w) &= 6\int_0^1 dw_{(3)} \int_0^{w_{(3)}} dw_{(2)} \int_0^{w_{(2)}} \delta\left(v_w - \frac{w_{(3)}}{w_{(1)}+w_{(2)}+w_{(3)}}\right) dw_{(1)}, \\ &= 6\int_0^1 dw_{(3)} \int_0^{w_{(3)}} \frac{w_{(3)}}{v_w^2} \mathbb{1}_{0<\frac{w_{(3)}-v_w\left(w_{(2)}+w_{(3)}\right)}{v_w}<w_{(2)}} dw_{(2)}, \end{aligned} \tag{S23}$$

or,

$$P_{v_w} = \begin{cases} 6\int_0^1 w_{(3)}^2 \frac{3v_w-1}{2v_w^3} dw_{(3)}, \text{ if } \frac{1}{3} < v_w \leq \frac{1}{2} & \text{(S24)} \\ 6\int_0^1 w_{(3)}^2 \frac{1-v_w}{2v_w^3} dw_{(3)}, \text{ if } \frac{1}{2} < v_w \leq 1 & \text{(S25)} \\ 0, \text{ otherwise.} & \text{(S26)} \end{cases}$$

Finally, after performing the integral, we get

$$P_{v_w} = \begin{cases} \frac{3v_w-1}{v_w^3} \text{ if } \frac{1}{3} < v_w \leq \frac{1}{2} & \text{(S27)} \\ \frac{1-v_w}{v_w^3}, \text{ if } \frac{1}{2} < v_w < 1 & \text{(S28)} \\ 0, \text{ otherwise.} & \text{(S29)} \end{cases}$$

2. Runner-up vote share distribution

Similarly, the probability density function of the runner-up vote share $v_r = V_r/T$ can be obtained as follows,

$$\begin{aligned} P_{v_r}(v_w) &= 6\int_0^1 dw_{(3)} \int_0^{w_{(3)}} dw_{(2)} \int_0^{w_{(2)}} \delta\left(v_r - \frac{w_{(2)}}{w_{(1)}+w_{(2)}+w_{(3)}}\right) dw_{(1)}, \\ &= 6\int_0^1 dw_{(3)} \int_0^{w_{(3)}} \frac{w_{(2)}}{v_r^2} \mathbb{1}_{0<(1/v_r-1)w_{(2)}-w_{(3)}<w_{(2)}} dw_{(2)}, \end{aligned} \tag{S30}$$

or,

$$P_{v_r}(v_r) = \begin{cases} 6\int_0^1 w_{(3)}^2 \frac{v_r(2-3v_r)}{2(1-v_r)^2(1-2v_r)^2} dw_{(3)}, \text{ if } 0 < v_r \leq \frac{1}{3} & \text{(S31)} \\ 6\int_0^1 w_{(3)}^2 \frac{1-2v_r}{2v_r^2(1-v_r)^2} dw_{(3)}, \text{ if } \frac{1}{3} < v_r < \frac{1}{2} & \text{(S32)} \\ 0, \text{ otherwise.} & \text{(S33)} \end{cases}$$

Finally, after performing the integral, we get

$$P_{v_r}(v_r) = \begin{cases} \dfrac{v_r(2-3v_r)}{(1-v_r)^2(1-2v_r)^2} \text{ if } 0 < v_r \leq \dfrac{1}{3} & \text{(S34)} \\ \dfrac{1-2v_r}{v_r^2(1-v_r)^2}, \text{ if } \dfrac{1}{3} < v_r < \dfrac{1}{3} & \text{(S35)} \\ 0, \text{ otherwise.} & \text{(S36)} \end{cases}$$

3. *Specific margin distribution*

We obtain the distribution of specific margin $\mu = M/T$ as follows,

$$\begin{aligned} P_\mu(\mu) &= 6\int_0^1 dw_{(3)} \int_0^{w_{(3)}} dw_{(2)} \int_0^{w_{(2)}} \delta\left(\mu - \frac{w_{(3)} - w_{(2)}}{w_{(1)} + w_{(2)} + w_{(3)}}\right) dw_{(1)}, \\ &= 6\int_0^1 dw_{(3)} \int_0^{w_{(3)}} \frac{w_{(3)} - w_{(2)}}{\mu^2} \mathbb{1}_{0 < \frac{w_{(3)} - w_{(2)} - \mu\left(w_{(3)} + w_{(2)}\right)}{\mu} < w_{(2)}} dw_{(2)}, \end{aligned} \tag{S37}$$

or,

$$P_\mu(\mu) = \begin{cases} 6\displaystyle\int_0^1 w_{(3)}^2 \frac{5 + 2\mu - 7\mu^2}{2(1+\mu)^2(1+2\mu)^2} dw_{(3)}, \text{ if } 0 < \mu < 1 & \text{(S38)} \\ 0, \text{ otherwise.} & \text{(S39)} \end{cases}$$

Finally, after performing the integral, we get

$$P_\mu(\mu) = \begin{cases} \dfrac{(1-\mu)(5+7\mu)}{(1+\mu)^2(1+2\mu)^2} \text{ if } 0 < \mu < 1 & \text{(S40)} \\ 0, \text{ otherwise.} & \text{(S41)} \end{cases}$$

D. Calculating the *scaled* distributions

The winner and runner-up vote shares and specific margins are random variables scaled by the voter turnout T. However, through a simple change of variable, $Y = yT$, we can obtain the conditional distributions of the unscaled variables as,

$$\mathcal{P}(Y|T) = \frac{1}{T} P_y(Y/T), \tag{S42}$$

where y can be v_w, v_r, and μ and Y represents unscaled variables V_w, V_r, and M respectively. The distribution of Y for an arbitrary turnout distribution $g(T)$ can be obtained as,

$$Q_Y(Y) = \int g(T)\ \mathcal{P}(Y|T) dT, \tag{S43}$$

with $\langle Y \rangle$ defined as,

$$\langle Y \rangle = \int Y\ Q_Y(Y) dY. \tag{S44}$$

Finally the distribution of *scaled* Y, defined as $\widetilde{Y} = Y/\langle Y \rangle$, can be obtained as follows,

$$Q_{\widetilde{Y}}(\widetilde{Y}) = \langle Y \rangle\ Q_Y(\widetilde{Y}\langle Y \rangle) \tag{S45}$$

Again, the dummy random variable Y can be either, V_w, V_r, and M.

E. The universal distribution of scaled specific margins

The distribution of scaled specific margins ($\widetilde{\mu} = \mu/\langle\mu\rangle$) is universal and can be obtained through a simple scaling of the distribution of specific margins from 3-candidate RVM by its mean. The distribution $Q_{\widetilde{\mu}}(\widetilde{\mu})$ has the following form:

$$Q_{\widetilde{\mu}}(\widetilde{\mu}) = \langle\mu\rangle \; Q_{\mu}(\widetilde{\mu}\langle\mu\rangle) = \frac{\langle\mu\rangle(1-\widetilde{\mu}\langle\mu\rangle)(5+7\;\widetilde{\mu}\langle\mu\rangle)}{(1+\widetilde{\mu}\langle\mu\rangle)^2(1+2\;\widetilde{\mu}\langle\mu\rangle)^2}, \tag{S46}$$

with $\langle\mu\rangle = \frac{1}{2} + \ln\left(\frac{9\sqrt[4]{3}}{16}\right)$.

S3. SIMULATION DETAILS

The Random Voting Model, $\mathcal{V}(T, n^c)$, with only voter turnouts and number of candidates as an input, can predict the distributions of the winner, runner-up votes, and the margins, when scaled appropriately. For the purpose of this model, the length of the voter turnout array can be assumed to be the total number of electoral units. The model can be simulated by drawing n^c random numbers from a uniform distribution $\mathcal{U}(0,1)$ for each electoral unit. These random numbers are further normalized using Eq. S1 to find the probabilities for attracting votes for each candidate. Next, each of the T_i (voter turnout in ith electoral unit) electors cast their votes according to the previously calculated probabilities. Finally, all the votes in that unit are counted, and votes received by the winner and the runner-up, along with the margin of victory, are stored. This is repeated for all the electoral units to obtain arrays of the winner votes, runner-up votes, and margins.

S4. PREDICTION OF SCALED MARGIN DISTRIBUTION AT DIFFERENT SCALES

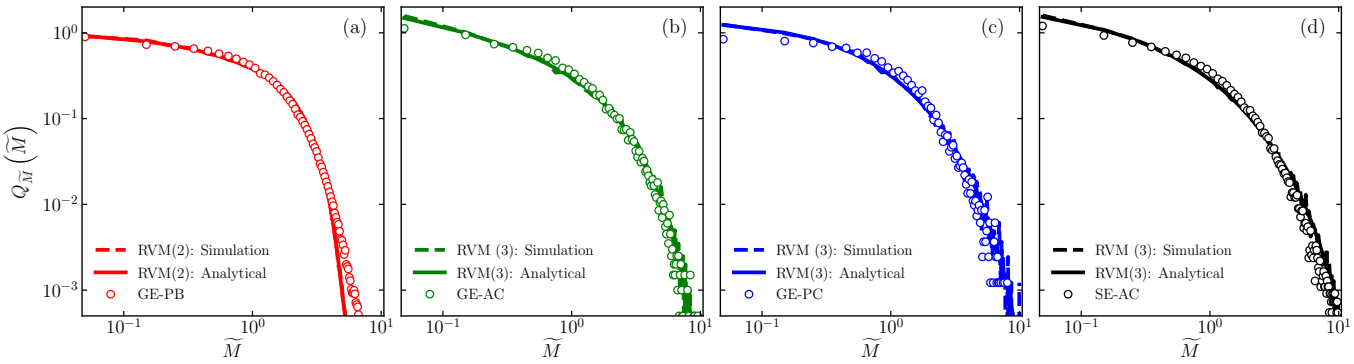

FIG. S1. Scaled distribution of margins at different electoral scales. Panel (a) - (c) demonstrates the excellent prediction of scaled margin distribution for Indian General Elections at the polling booth, assembly constituency, and parliamentary constituency levels, respectively. Panel (d) demonstrates the same for state Assembly Elections at the assembly constituency level. The open circles denote the empirical distributions. The dashed and solid lines denote the prediction from RVM simulations and analytical calculations, respectively.

S5. DATA SUMMARY

The parliamentary and assembly constituency level data of Indian General elections, along with the assembly constituency level data of the Assembly election, were obtained from the election data repository of Lok Dhaba [41]. The polling booth level data for the general elections were collected from the websites of chief electoral officers of different states in India [40]. The following table contains the summary of Indian election data.

Election Type	General Election	General Election	General Election	State Election
Electoral Scale	Parliamentary Constituency	Assembly Constituency	Polling Booth	Assembly Constituency
Time Span	1962-2019	1999-2019	2004-2019	1961-2023
Number of Elections	52 (including bye-elections)	5	4	61
Average Turnout	587329	116577	583	86484
Average Winner Votes	286807	56874	348	39884
Average Runner-up Votes	201281	38887	159	28562
Average Margin	85526	17987	189	11322

TABLE S1. The table presents typical values of voter turnout and winner votes, runner-up votes, and winning margins across different electoral levels for various types of Indian elections. The available data for the mentioned time spans were consolidated for each country and used to calculate the respective averages. The data for an electoral unit is considered valid if it meets the following criteria: (a) a complete list of votes received by all candidates was available, (b) at least two candidates contested the election, and (c) the turnout was non-zero.