http://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 10, No. 3, September 2020, pp. 193–222
ISSN 1946-5238 (online)

Crowdvoting Judgment: An Analysis of Modern Peer Review

Michael R. Wagner[a]

[a] Michael G. Foster School of Business, University of Washington, Seattle, Washington 98195
Contact: mrwagner@uw.edu, https://orcid.org/0000-0003-2077-3564 (MRW)

Received: January 3, 2018
Revised: July 2, 2018; March 31, 2019
Accepted: July 28, 2019
Published Online in Articles in Advance: June 2, 2020

https://doi.org/10.1287/stsy.2019.0053

Abstract. In this paper, we propose and analyze models of self policing in online communities, in which assessment activities, typically handled by firm employees, are shifted to the "crowd." Our underlying objective is to maximize firm value by maintaining the quality of the online community to prevent attrition, which, given a parsimonious model of voter participation, we show can be achieved by efficiently utilizing the crowd of volunteer voters. To do so, we focus on minimizing the number of voters needed for each assessment, subject to service-level constraints, which depends on a voting aggregation rule. We focus our attention on classes of voting aggregators that are simple, interpretable, and implementable, which increases the chance of adoption in practice. We consider static and dynamic variants of simple majority-rule voting, with which each vote is treated equally. We also study static and dynamic variants of a more sophisticated voting rule that allows more accurate voters to have a larger influence in determining the aggregate decision. We consider both independent and correlated voters and show that correlation is detrimental to performance. Finally, we take a system view and characterize the limit of a costless crowdvoting system that relies solely on volunteer voters. If this limit does not satisfy target service levels, then costly firm employees are needed to supplement the crowd.

Funding: M. R. Wagner gratefully acknowledges the support of a Neal and Jan Dempsey Faculty Fellowship.

Keywords: crowdsourcing • crowdvoting • online communities • random walks • sequential hypothesis testing

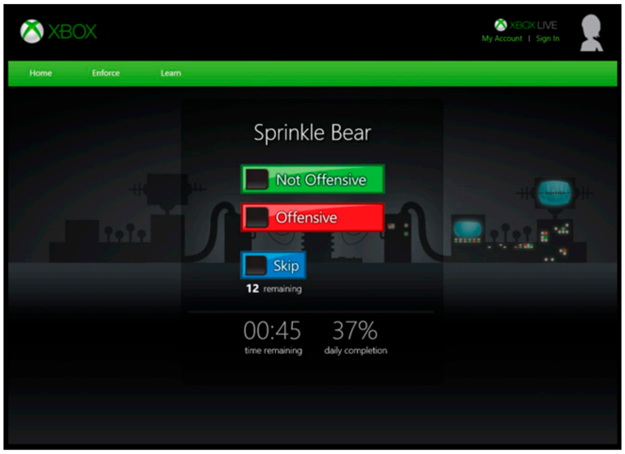
1. Introduction

A unique feature of the modern video game industry is the ability of gamers to play together online, which results in large online communities of gamers: Burns (2013) reports that Microsoft's Xbox Live service has more than 48 million users, the Steam gaming network has 65 million users, and the Playstation network has 110 million users. These online communities typically have terms of service (ToS) that prescribe acceptable online behavior, which can be violated, and thus, policing is required to preserve the value of the community/firm. Firms can assign internal employees to handle this policing in-house, or they can outsource it; Chen (2014) reports on a workforce of more than 100,000 in the Philippines that provide "content moderation" for online communities.

In contrast to employee-led or outsourced policing, some firms have recently started to utilize the online community itself (i.e., the crowd) in judging offensive behavior. In August of 2013, the Xbox Live service announced its "Enforcement United" program (enforcement.xbox.com/United), by which gamers vote on whether flagged content violates the Xbox Live ToS. League of Legends is another online gaming platform with more than 32 million active monthly users (MacManus 2012), that utilizes self-policing online communities via its Tribunal program (leagueoflegends.com/tribunal). Note that, as of this writing, we are not aware of any crowdvoting applications on the Steam and Playstation networks.

In Figure 1, we provide a screenshot of a sample assessment in the Enforcement United program. This assessment stems from a complaint about an in-game alias "Sprinkle Bear" selected by one user, which another user found offensive. A distinct set of gamers are randomly selected as a jury for this assessment, and they each individually see the view in Figure 1. Each voter must click either the green "Not Offensive" button if they believe "Sprinkle Bear" is acceptable in the Xbox Live ToS or the red "Offensive" button otherwise. These assessments have a short time limit (e.g., one minute) and can clearly be done quickly. This structure results in minimal effort on the part of the voter, which allows multiple assessments to be performed in quick

Figure 1. Screenshot of Enforcement United



succession by a given voter; this observation motivates our model of voter participation, detailed in Section 2.1. Finally, a voter is allowed to press the blue “Skip” button, skipping a given assessment; for simplicity, we omit this option from our analysis.

In this paper, we introduce and study models of self-policing systems that are motivated by these innovative applications. We first take the firm’s perspective of maximizing firm value subject to minimum service-level constraints. This view leads us to the notion of efficient crowdvoting, by which we identify the voting mechanism(s) that require the fewest voters per assessment. More specifically, we propose and analyze numerous voting mechanisms, both static and dynamic, and determine the probability that the crowd is able to correctly determine the underlying truth, which we relate to the service level; these mechanisms are rigorously developed using ideas from basic probability theory, random walks, and sequential hypothesis testing. Furthermore, we limit our attention to classes of voting mechanisms that are simple, interpretable, and implementable; by doing so, we increase the likelihood of adoption in practice and creating value for a firm.

These voting mechanisms are then fed into optimization models in which the number of voters per assessment is minimized (which leads to maximized firm value), for each mechanism, subject to bounds on error probabilities. Our paper effectively shows that, for self-policing online communities, small crowds (of juries) are best.

1.1. Broad Applicability

Crowdvoting, the focus of our paper, utilizes the crowd’s judgment to evaluate content. There are many potential applications beyond the gaming context described, which we now discuss. In these crowdvoting examples, users judge other users, which is a form of peer review.

Twitter and Facebook are two of the largest social media platforms. As of this writing, Twitter has 321 million active monthly users (https://www.statista.com/). Similarly, Facebook currently has more than 2.3 *billion* active monthly users (newsroom.fb.com/company-info/). As in the video game communities, there are ToS-violating behaviors on these social media online communities that result in complaints. For example, O’Brien (2016) reports that Facebook receives 1 million user violation reports per day. There is also evidence that these firms are not doing a good job at content moderation: Tiku and Newton (2015) report that Twitter’s CEO stated, “We suck at dealing with abuse.” To the best of our knowledge, neither of these firms use crowdvoting to moderate content; indeed, Facebook is hiring more employees to screen offensive content (Goel 2017). Perhaps self policing, as described in our paper, would be a better answer. A recent *Washington Post* article (Dwoskin 2018) describes how Facebook is rating users on their trustworthiness on a zero to one scale; therefore, applying the results in this paper would be a simple matter for Facebook.

A related potential application of the crowdvoting models studied in this paper is the detection of “fake news,” intentionally false and misleading news articles, typically disseminated via social media platforms, such as Twitter and Facebook. Recent studies suggest that fake news influenced the outcome of the 2016 U.S. presidential election (Blake 2018). Facebook, for instance, recognizes the deleterious influence of fake news and is attempting to reduce its impact (Thompson 2018). To the best of our knowledge, there is little academic study of identifying fake news; one exception is Papanastasiou (2020), which studies the platform’s optimal inspection policy. Crowdvoting, as studied in our paper, provides another potential solution.

There are similar applications of crowdvoting in product development. Threadless (https://www.threadless.com/designs/) is a firm that allows the crowd to vote on t-shirt designs and then manufactures the most popular ones. Instead of evaluating whether a complaint is a violation of ToS, in this example, crowdvoting is used to evaluate whether a product will be successful or not. Similarly, PepsiCo utilized crowdvoting in multiple campaigns to determine new flavors of its Mountain Dew soft drink: (a) in 2007–2008, DEWmocracy resulted in the introduction of the new Voltage flavor, (b) in 2009–2010, DEWmocracy II resulted in the new WhiteOut flavor, and (c) in 2013, DEWmocracy Canada resulted in Voltage being declared the winner (outperforming White Out). Note that there can also be overlap between crowdvoting and other types of crowdsourcing: for example, in addition to submitting ideas on Dell's IdeaStorm website (i.e., ideation), users can also vote (promote/demote) on other users' ideas.

Social news websites also utilize crowdvoting. Digg.com and Reddit.com are two news aggregators, on which users submit and vote (promote/demote) on news articles, and the more popular ones are displayed more prominently on the websites. A similar application of crowdvoting even appears in the field of academic writing: http://tex.stackexchange.com is an online forum on which any user can pose questions about Latex, any user can answer the questions, and any user can vote (promote/demote) a question or answer. The more popular combinations are then displayed more prominently on the website. Furthermore, http://stackexchange.com/sites is an aggregator that lists many similar sites on varied topics, such as computer programming, photography, languages, personal finance, and (even) homebrewing. Thus, similar to the product-development examples, crowdvoting is currently being used to assess the popularity of news articles, questions, and answers.

It is also possible that other prominent crowdsourcing projects could benefit from crowdvoting. For instance, the development of the operating system Linux is perhaps one of the most successful examples of crowdsourcing. However, new modifications/additions to Linux must be approved by a small management team led by Linux founder Linus Torvalds himself; see the Linux Information Project at http://www.linfo.org/ for further details. Perhaps this management team should consider crowdvoting the approval process using the results in our paper; after all, if the crowd is capable of improving Linux, perhaps they are also capable of approving modifications.

Finally, our results could also be applied in the context of academic peer review. After an academic paper is submitted to a journal for possible publication, a small number of referees and editors are recruited (typically without pay) to assess the appropriateness of a submission. The results in this paper could be used to determine a proper number of referees as well as to more appropriately and formally aggregate each reviewer's vote on the suitability of the paper for publication. Detailed examples of such an application are provided in the sequel.

1.2. Literature Review

Our crowdvoting research is related to political science, economics, and crowdsourcing research, and we position our work with respect to each of these areas.

A classic result of political science is Condorcet's jury theorem, which states that, if n voters each independently choose a correct alternative (out of two choices) with a known probability $p > 0.5$, then the probability that a majority vote is correct is greater than p and converges to one as $n \rightarrow \infty$. This famous theorem was generalized to correlated votes with voter-specific accuracies p_i by Boland (1989) and Ladha (1992), and Berend and Paroush (1998) established precise necessary and sufficient conditions for the jury theorem to hold, by which, in all cases, the p_i are known constants. Some of our basic results could be presented as variants of Condorcet's jury theorem except that, in our setting, voter accuracies are independent and identically distributed (i.i.d.) random variables rather than a deterministic sequence as in the literature. However, because the number of voters can be controlled in a crowdvoting application (as opposed to a political election), our focus is instead on studying the transient behavior of voting rules for finite n, providing simple formulae that can help manage a crowdvoting system. In addition, we also consider voting rules under which the number of voters is not specified in advance and is determined dynamically as voters arrive to the system.

There is a substantial literature studying group decision making in economics, and here we cite the references most relevant to our paper. Sah and Stiglitz (1986) study hierarchies and polyarchies, and Sah and Stiglitz (1988) primarily analyze committees in the context of selecting projects that result in positive or negative profits (i.e., good or bad projects). In particular, Sah and Stiglitz (1988) derive optimal committee sizes and consensus levels to maximize expected profit. Ben-Yashar and Nitzan (1997) study a similar problem and derive optimal voting weights, again to maximize expected profit. Our basic model is different than these papers in that we have no similar notion of a project and project cash flows because the crowd is voting on an *assessment*, which does not have an immediate profit or loss. Instead, our basic model is to make the most cost-effective use of expensive (infallible) employee voters *and* costless volunteer voters, whose capacity for voting is limited; these two classes of voters, with different costs, have not been studied in the literature to the best of our knowledge, and this is a unique characteristic of crowdvoting. Therefore, our paper complements this literature stream. Furthermore, the voters' accuracies in our paper are random variables, whereas, in these papers, the corresponding accuracies are known constants. In another stream of economic work, group decision making is analyzed via game theory. Feddersen and Pesendorfer (1997) analyze a game in which each voter has a known utility function and characterize the Nash equilibrium under various preferences and informational environments. Feddersen and Pesendorfer (1998) similarly characterize the Nash equilibrium of a jury, demonstrating that, if voters act strategically, requiring unanimity in the votes is inferior to simple majority voting in the sense that the probability of convicting an innocent person (or acquitting a guilty person) is higher; Ladha et al. (1996) find empirical support for strategic votes in a jury. A game theoretic formulation is not applicable to our context because voters in a crowdvoting system do not interact in contrast to political (jury) voters who usually (must) interact. Furthermore, the details of the crowdvoting mechanisms are typically hidden from voters, which minimizes any motivation for system gaming and strategic actions. For further references, Gerling et al. (2005) and Li and Suen (2009) are related surveys.

We next summarize the related crowdsourcing literature. Small, quick jobs ill-suited for a computer (e.g., identifying the subject of a photograph) are usually called human intelligence tasks (HITs), and Amazon's Mechanical Turk system (mturk.com) is perhaps the best-known platform for crowdsourcing HITs. A study of the Mechanical Turk system, closely related to our paper, is Amir et al. (2013), which analyzes the probability of correct crowd assessment; these authors show, via human experiments, that the crowd has a higher probability of correct crowd assessment for NP-complete problems (hard to solve, easy to verify) than PSPACE-complete (hard to solve, hard to verify) problems. We, alternatively, take an analytical approach that complements this experimental approach. Acemoglu et al. (2019) formulate and analyze a dynamic programming model for resource allocation in a crowd with unobservable skills for an HIT context. Karger et al. (2014) utilize a similar model to ours in the context of a crowdsourcing system for HITs, in which they also model an unobservable truth, and workers' reliabilities are random variables. These authors study the problem of minimizing the cost of utilizing workers to achieve a target overall reliability, and they design an algorithm for assigning tasks to workers; in contrast, in our paper, the crowd members are volunteers and are not paid financially; furthermore, our solutions are closed form (not algorithmic), which results in increased interpretability and ease of implementation. Liu et al. (2014) perform a field experiment on Taskcn (taskcn.com), an online Chinese labor market similar to Amazon's Mechanical Turk system, and study the effect of reward levels and existing submissions (which are visible to all by default). The theoretical foundation of this paper consists of a model that has some similarities to our own: users' abilities are modeled as random variables on the interval $[0,1]$ (which is also similar to Karger et al. (2014)). Again, our analytical results complement the experimental emphasis of this paper. The main difference between this stream of work on HITs and our paper is that, in an HIT application, the crowd members are *paid* for their work, whereas the crowdvoting application studied in our paper is based on *volunteer* voters, which results in a fundamentally different model. Furthermore, in HIT applications, the crowd workforce is transient, whereas, in our crowdvoting applications, the crowd is stable as it typically consists of an online customer base (e.g., Xbox Live). Massoulié and Xu (2018) study a problem that is similar and more general to ours, in that agents of multiple types are tasked with evaluating content and provide noisy assessments, and the authors develop an algorithmic solution that is asymptotically optimal; although their model is general, their algorithmic solution is not transparent or interpretable (e.g., an optimization model must be solved numerically in a subroutine), resulting in a black-box solution. In contrast, our analysis is finite (i.e., not asymptotic) and our results are closed form, interpretable, and easily implementable, which results in a solution that is arguably more likely to be adopted in practice.

Budescu and Chen (2015) study the wisdom of the crowd and design an algorithm to identify the poorly performing individuals who are excluded from the crowd; we similarly identify a simple modification that allows us to neutralize the weakest voters (but this is not a major consideration in our paper). Papanastasiou et al. (2018) study a platform's information control policy to influence the crowd to take consumer-surplus optimal actions; our basic problem is different in that our fundamental decision is not to control information but rather to decide what size crowd is needed per assessment. Marinesi and Girotra (2013) formally study customer voting systems in the context of product development and pricing; they show that when customers vote on product *development*, firm and customer interests are aligned, whereas, when customers vote on product *pricing*, the voting systems are ineffective because of misaligned objectives. Similarly, Caldentey and Araman (2013) study the use of crowdvoting to determine the timing of new product introductions.

1.3. Our Contributions

In this paper, we present analytical models of crowdvoting in self-policing online communities, motivated by recent innovative applications in the video game industry, in which members of the service, not employees, evaluate complaints. Our overarching aim is to introduce parsimonious models that result in simple and interpretable crowdvoting rules to efficiently utilize the crowd that are easy to implement so that the likelihood of adoption in practice is high. Our results can potentially be applied to similar self-policing applications at Facebook, Twitter, and other social media platforms; detection of fake news on these platforms; identifying products that have high potential for success; identifying popular topics on news aggregators or question–answer websites; and academic peer review.

Although many voting models exist in the literature, our approach is unique for two reasons. First, our models are not driven by project selection with stochastic cash flows, a common objective in the (economics) literature; rather, in a crowdvoting system, voters evaluate an assessment that has no immediate cash flows associated with it. However, our research is motivated by maximizing firm value, which we argue leads to a model of using the costless volunteer voters most effectively, resulting in the cost minimization of expensive firm experts. Second, we model voter accuracies as random variables rather than known constants, a modeling choice motivated by the random selection of voters from a large crowd; this modeling aspect is common in the crowdsourcing literature but rare in other voting literatures.

Our paper is also fundamentally different than much of the crowdsourcing literature, especially that on HITs, with which the crowd participants are paid a wage and constitute a transient workforce; instead, our crowdvoting framework models volunteer voters who are not paid and exploits the stable nature of the online community to learn and utilize the heterogenous abilities of the different members. Furthermore, we optimize over classes of voting mechanisms that are simple, interpretable, and implementable (rather than algorithmic), which increases the chance of adoption in practice and value creation for a firm.

The underlying premise of our paper is the objective of maximizing firm value. We argue that, using the unique characteristics of crowdvoting, minimizing the number of volunteer crowd voters per assessment, subject to service-level constraints, effectively minimizes the cost of policing the online community, thus maximizing firm value. To solve this minimization problem, we analyze the probability that a randomly selected subset of n voters from the crowd correctly determines the unobservable truth for a given assessment, for a variety of static and dynamic voting mechanisms that are both interpretable and implementable. Regarding the static rules, we obtain closed-form expressions for this probability, for simple majority-rule voting as well as a variant in which more accurate voters get a more heavily weighted vote under independent voters. The introduction of correlated voters complicates the analysis, which we resolve using numerical experiments; we find that correlated voters are detrimental to system performance, which is due to diminished crowd diversity, a characteristic that the literature has shown is beneficial for crowdsourcing applications (e.g., see Terwiesch and Xu 2008). We also allow the number of voters to be determined dynamically as voters arrive to the system rather than in an a priori fashion. In majority-rule voting, we utilize random walks, and in accuracy-weighted voting, we utilize sequential hypothesis testing. We obtain closed-form expressions for the probability that the crowd correctly determines the unobservable truth as well as the expected number of voters under independent voters.

In general, the accuracy-weighted rules outperform the simpler majority rules. However, these strong performances come at a price: even when voter correlation is not present, both accuracy-weighted rules can be difficult to implement in practice because of more intensive parameter estimation requirements. This performance-implementation trade-off provides a rationale for the adoption of the simpler majority-voting rules in practice. This trade-off is discussed in depth in our paper.

The final part of our paper considers the stream of assessments that arise in a practical application of crowdvoting. Building on the idea that a single user will likely vote on a few, but not many, assessments in a short period of time, we consider the matching of the supply of voters with the demand for assessments. Introducing the rate at which a population generates complaints, we show that there is an intrinsic limit to the crowd's ability to assess content if the complaint rate is too high; stated differently, we are able to characterize the limits of a *costless* assessment system that depends solely on the crowd. The managerial implication is that the crowd might need to be complemented by firm employees to attain target service levels.

All proofs are presented in Appendix A.

2. The Firm's Problem

A key concept of our paper is *crowdvoting*, which we define as the set of activities in which a firm uses a collection of volunteer persons, not employed by the firm, to evaluate content that has an objective value. We assume the following:

Assumption 1. *A firm has a limited pool of costly experts that can perfectly evaluate content.*

Assumption 2. *The firm has an associated online community (crowd) that is a large though not unlimited resource of nonfirm persons that imperfectly evaluate content.*

We begin with the premise that the firm is interested in maximizing its value. A potential source of value reduction is misbehaving users that violate the ToS, which reduces the attraction of the community to other users and could lead to customer attrition and lower firm revenues. Therefore, firms have traditionally used internal or external employees to perform policing functions to preserve the quality of the online community (i.e., adherence to the ToS). However, a potential source of cost savings is to replace some or all of the costly employees with members of the online community who are willing to assess complaints without financial compensation. In our paper, we focus on the cost-minimization perspective of the firm. Conceptually, we consider the problem

$$\begin{aligned} \min \quad & \text{(Crowd Assessment Cost)} + \text{(Employee Assessment Cost)} \\ \text{s.t.} \quad & \text{Service Constraints.} \end{aligned} \tag{1}$$

The service constraints contain three constraints. The first is to ensure that the decision reached by management is correct with high probability; for the employee assessors, this is assumed to be one per Assumption 1; for the crowd assessments, we devote a large portion of this paper to calculating the probability of correct crowd assessment (PCCA) for various voting mechanisms, which is bounded from below in problem (1). The other two constraints bounded from above are the probability of type I and II errors; again, these probabilities are zero for the employee assessors, and for the crowd, they follow from the PCCA calculations. Therefore, the service constraints exclusively pertain to the crowd assessments.

The objective consists of two components, the first of which is the crowd assessment cost. Note that the crowdvoting examples mentioned in the introduction depend on *volunteer voters*, who are not compensated financially. In other words, the crowd assessment cost is zero. The second cost, employee assessment cost, can be derived from the wage rate of the employees; however, this specific calculation is not needed because we simply need to assign as many assessments to the costless crowd as possible while still satisfying the constraints, and any remaining assessments get assigned to the employees.

Although large, the online communities are not unlimited, and it is not always possible to assign all assessments to the crowd while satisfying the service constraints. This is due to voter participation or lack thereof, which we next discuss.

2.1. Voter Participation

The crowdvoting examples mentioned in the introduction depend on volunteer voters, who are not compensated financially. Although a single voter could potentially participate even without financial compensation, the voter's participation is clearly not unlimited. To capture this intuitive effect, we adopt and modify the Downs (1957) model of voter participation as presented by Riker and Ordeshook (1968). This model states that a single voter's net return r from voting can be written as $r = pb + d - c$, where p is the probability that the vote is decisive, b is the voter benefit when the vote is decisive, d is the positive benefit of voting not associated with the outcome, and c is the positive cost of voting not associated with the outcome; the voter votes if and only if $r > 0$. In our crowdvoting context, we propose that a *single* voter's return r_i on a *single* assessment i is

$$r_i = u_i - c_i,$$

where $u_i = p_i b_i + d_i$ is the overall utility of voting on assessment i, $b_i \geq 0$ measures the value of influencing the outcome of an assessment, $p_i \geq 0$ is the probability that the vote is decisive in the assessment, $d_i \geq 0$ measures voter satisfaction, and $c_i > 0$ is the time commitment required to make the assessment. Because knowing whether a vote is decisive or not is typically unobservable, we focus on discussing the voter satisfaction d_i. Consider open source software, for which software development is primarily accomplished by volunteers who receive only nonfinancial compensation, such as pride; d_i can capture this pride. Microsoft's Xbox Live service has an "Ambassadors" program (https://ambassadors.xbox.com), by which volunteers assist gamers and are only compensated in "levels and loot," which are rewards redeemable only within the Xbox Live ecosystem; d_i can capture a user's valuation of these nonfinancial offerings. Anderson et al. (2013) analyze the incentives that result from nonfinancial *badges*, which are awarded to users for various achievements and appear on a number of websites to increase user participation (e.g., Huffington Post, Khan Academy, and Wikipedia); they show that these badges can be powerful incentives that increase user participation; d_i can capture the value of these badges.

In practice, a single voter is prompted to perform many assessments sequentially but can stop at any time. For example, when a gamer logs into the Xbox Live service, there is a prompt asking the gamer to perform a number of assessments; the prompt is a request, and the gamer has no obligation to do the assessments. It is reasonable to assume that $c_i = c$ for all i; as discussed, this cost is typically the time needed to perform each assessment, which is usually one minute in the Enforcement United case. We also assume that u_i is decreasing in i, where i indexes the ith assessment provided to a gamer, to reflect the fact that a gamer generally loses interest as the gamer performs more assessments in a row. If a voter decides to perform m assessments and then quits to play video games, the voter's cumulative return is $\sum_{i=1}^{m} r_i = \sum_{i=1}^{m}(u_i - c)$. The number of assessments m is chosen to maximize the voter's cumulative return, $\max_m \sum_{i=1}^{m} u_i - cm$, which clearly has a unique maximum m^* whenever u_i strictly decreases to zero in i. This model reflects the reality that a voter is likely to perform, say, five assessments in a row taking five minutes, but is unlikely to perform 100 assessments in a row.

This discussion provides evidence that crowd votes are a limited resource that the firm must utilize efficiently. One way to do so is to minimize the number of voters per assessment, which maximizes the number of assessments assigned to the costless crowd. Equivalently, this minimizes the number of assessments assigned to the costly firm employees, thus minimizing firm cost. Consequently, we indirectly solve problem (1) by solving the following constrained mathematical programming model:

$$\begin{aligned} \min \quad & \text{Number of Voters per Assessment} \\ \text{s.t.} \quad & \text{Probability of Correct Crowd Assessment} \geq \gamma \\ & \text{Probability of Type I Error} \leq \epsilon \\ & \text{Probability of Type II Error} \leq \varepsilon, \end{aligned} \tag{2}$$

where γ, ϵ, and ε are user-supplied probabilistic thresholds that capture service-level constraints. In particular, our objective is to optimize over simple, interpretable, and easily implementable classes of voting mechanisms (described in Section 3.2); the focus on easy to understand and implement voting mechanisms effectively maximizes the likelihood of adoption in practice and value creation for firms. Problem (2) becomes our canonical model, which we solve for various interpretable voting mechanisms in subsequent sections, in which full details of the optimization models appear.

3. Crowdvoting Models

An important primitive of our model is an assessment, such as whether a user has violated the ToS of an online community. A user's complaint typically generates the need for an assessment (e.g., a participant notices that another participant's username is offensive). Management is then responsible for assessing the complaint and determining whether further action is necessary (e.g., removing the participant from the community). Historically, management used employees (firm experts) to assess the complaints, but recently firms have been utilizing crowdvoting to outsource the process to the crowd.

We primarily consider binary assessments and let T denote the truth underlying a given assessment, where $T \in \{-1, 1\}$; in Section 7, we discuss assessments with more than two outcomes. If $T = 1$, we associate a positive connotation (e.g., a user has *not* violated ToS), and for $T = -1$, we associate the complementary, negative connotation (e.g., a user has violated ToS). As discussed in the previous section, our assumptions imply that the firm, using employee experts, has the ability to ascertain the value of T with certainty. However, the firm might not have a sufficient number of experts to evaluate all complaints internally; recall that the Xbox Live community has at least 48 million users, and every member is a potential complaint generator. Alternatively, the firm simply might not want to assess all complaints internally because of cost considerations, which further motivates problem (1). Therefore, the firm outsources some or all of the assessments to a crowd of *imperfect* voters who are members of the firm's associated online user community, and any remaining assessments are evaluated by internal firm experts.

3.1. Stochastic Accuracies for Crowd Voters

If a given assessment is assigned to the crowd, we assume that n voters are selected randomly from the crowd to vote on the assessment; we further assume that n is much smaller than the size of the crowd. There is typically heterogeneity in voters' abilities to ascertain the underlying truth, which reflects differing voter skill, bias, and even ignorance. We let voter i's vote $v_i \in \{-1, 1\}$ for $i = 1, \ldots, n$ and let $p_i \in [0, 1]$ denote the accuracy of voter i in identifying $T = 1$; mathematically, $P(v_i = 1|T = 1) = p_i$. Similarly, we let $q_i \in [0, 1]$ denote the accuracy of voter i in identifying $T = -1$; mathematically, $P(v_i = -1|T = -1) = q_i$.

The p_i and q_i accuracies might or might not be observable (i.e., estimated with a reasonable degree of precision). If voters participate frequently, the system can estimate the p_i and q_i values for voter i using historical data. For example, k historical assessments indexed by $j = 1, \ldots, k$ with known values of $T_j = 1$ (determined by firm experts), can be given to voter i, and an estimate for p_i can be determined by how many times voter i agrees with the known underlying truth; mathematically, the estimate $\hat{p}_i = \sum_{j=1}^{k} 1\{v_j = 1\}/k$ can be used as a proxy for p_i, where $1\{\}$ is the indicator function. Of course, there is estimation error. Standard statistical theory, under mild conditions, prescribes $\left[\hat{p}_i - z\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{k}}, \hat{p}_i + z\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{k}}\right]$ as a $(1-\alpha)\%$ confidence interval for p_i, where z is the $(1-\alpha/2)$ percentile of a standard normal distribution. Although we could increase k to have a more precise estimate of p_i, this is taxing on a single voter; for instance, if we desire the estimate $\hat{p}_i$ to be within, say, 1% of the true value p_i with, say, 99% confidence, then, noting that $\hat{p}_i(1-\hat{p}_i) \le \frac{1}{4}$, the confidence interval can be manipulated to show that $k \ge \frac{(\Phi^{-1}(0.995))^2}{4(0.01)^2} \approx 16{,}587$ (training) assessments are needed. Alternatively, we suggest to instead tax the crowd by choosing a relatively small value of k and letting the lower limit of the confidence interval, $\hat{p}_i - z\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{k}}$, be the proxy for p_i. For example, if $k = 100$, $z = \Phi^{-1}(0.995) = 2.576$, and, say, $\hat{p}_i = 0.80$, the lower bound equals $\hat{p}_i - z\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{k}} = 0.70$. In this way, the voter's accuracy is underestimated, and more voters are needed for every assessment. However, each voter would be available to start contributing to nontraining crowdvoting much sooner, providing a larger voting population to draw from, which effectively neutralizes the need for more voters per assessment.

In the sequel, for analytical convenience, we assume that the p_i and q_i values are known for all (participating) members of the crowd though we also provide voting mechanisms that do not require these values in case it is burdensome to estimate them. However, for a given assessment, because a *random* sample of n voters is selected from the crowd, a priori the values of (p_i, q_i), $i = 1, \ldots, n$, for the assessment are not known. We, therefore, introduce distributions for p_i and q_i to reflect this random selection of a subset of voters from the crowd for a given assessment:

Assumption 3. *The voter accuracies p_i and q_i, $i = 1, \ldots, n$, for a given assessment, are random variables:*

- *Voter accuracies p_i in identifying $T = 1$ are draws from a continuous probability distribution f_p with support on $[0, 1]$ with mean μ_p, standard deviation σ_p, and pairwise correlation coefficient r_p.*
- *Voter accuracies q_i in identifying $T = -1$ are draws from a continuous probability distribution f_q with support on $[0, 1]$ with mean μ_q, standard deviation σ_q, and pairwise correlation coefficient r_q.*

Assuming the crowd has size N, the crowd data, (p_i, q_i), $i = 1, \ldots, N$, can be used to determine an empirical estimation of the distributions f_p and f_q as well as estimate the statistics μ_p, μ_q, σ_p, σ_q, r_p, and r_q. Alternatively, the crowd data can be used to fit a well-known distribution; in our opinion, a good choice is the beta distribution, a common conjugate prior in Bayesian statistics for a probability parameter. In addition, our subsequent results are robust with respect to the support of these distributions. For example, if we only utilize voters whose accuracies p_i are contained in the interval $[a, b] \subset [0, 1]$, then the conditional distribution $f_p(t)/\int_a^b f_p(\tau)d\tau$, $t \in [a, b]$, can be used in place of f_p in all our results. If historical assessments are not available to determine the crowd data (p_i, q_i), $i = 1, \ldots, N$, then a uniform distribution, potentially with support $[a, b] \subset [0, 1]$, can be used for f_p and f_q; this is an appealing option because the uniform distribution is entropy maximizing. Although our core results assume that the distributions are static, in Section 7.1, we discuss the potential for voter learning (i.e., improving accuracies over time and changing distributions). Finally, as mentioned, we also provide voting mechanisms that require minimum information about the accuracy distributions in case it is burdensome to estimate them in practice.

Similarly, we assume that there is a prior distribution for the underlying truth.

Assumption 4. *The underlying truth T is a Rademacher random variable with parameter ϱ:*

$$P(T = 1) = \varrho \qquad \text{and} \qquad P(T = -1) = 1 - \varrho.$$

As discussed, the parameter ϱ can also be determined by looking at the history of past assessments. This assumption allows us to introduce the unconditional accuracy of voter i:

Definition 1. Voter i's unconditional accuracy is defined as ρ_i, which is modeled as a mixture of the p_i and q_i random variables:

$$\rho_i = \begin{cases} p_i, & \text{with probability } \varrho \\ q_i, & \text{with probability } 1 - \varrho. \end{cases} \tag{3}$$

Note that, although correlation between p_i and q_i might (and probably does) exist, it is not relevant for our analysis; for a given assessment, the underlying truth has a specific value (say, $T = 1$) and only one conditional accuracy is relevant (e.g., p_i for $T = 1$). In other words, the random variables p_i and q_i are never combined and used concurrently in a given assessment. Furthermore, note that the distribution of ρ_i can be readily calculated from the distributions f_p and f_q: $f_\rho = \varrho f_p + (1 - \varrho) f_q$. However, the correlation between p_i and p_j (or q_i and q_j), $i \neq j$, equals r_p (r_q) and factors into our analysis when it exists.

3.2. Voting Rules

The vector $v = (v_1, \ldots, v_n)$ represents all vote values. We propose the following decision rule for the crowdvoting system:

$$R(v) = \begin{cases} 1, & \sum_{i=1}^{n} w_i v_i \geq m_p \\ -1, & \sum_{i=1}^{n} w_i v_i \leq -m_q \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $w_1, \ldots, w_n$ are nonnegative weights. If $R(v) = 1$, the system concludes that $T = 1$; if $R(v) = -1$, the system concludes that $T = -1$; and if $R(v) = 0$, the crowd is inconclusive and the system uses a (costly) firm expert to make the assessment. The parameters m_p and m_q allow a decision maker to select consensus levels. We only assume that $-m_q \leq m_p$ but otherwise make no assumptions on the sign of these parameters; we do, however, subsequently make recommendations on how to set m_p and m_q.

We consider two sets of weights in this paper: (a) unit weights $w_i = 1, \forall i$ and (b) accuracy weights $w_i = \rho_i, \forall i$, for which a more accurate voter casts a more heavily weighted vote. When unit weights are applied, our decision rule is the standard majority voting rule. When accuracy weights are utilized, our rule is analogous to the methodology of Nate Silver's website FiveThirtyEight.com (a polling aggregator), which weights each pollster by accuracy in its aggregation; see Felder (2009) for further details of the weighting methodology. This website correctly predicted the winner of all 50 states and the District of Columbia in the 2012 presidential election (Salant and Curtis 2012). Note that, once voter i is selected for an assessment, p_i and q_i are known. The weight ρ_i, in contrast, is a random weight, according to Equation (3), which is easily implemented in practice with a known value of ϱ.

The literature also identifies the log-odds weights $w_i = \log(\rho_i/(1 - \rho_i))$, which maximize the probability that the crowd makes the correct assessment for a given n; see Grofman et al. (1983) for a proof when the ρ_i are deterministic. Unfortunately, we were unable to obtain our closed-form solutions for these weights. However, we demonstrate in Monte Carlo simulation studies that the suboptimality of our simpler accuracy weights ($w_i = \rho_i$) is minimal (less than 1% loss on average); these results appear in Appendix B.

Before discussing our analysis and main results, we present notation for the reader's convenience in Table 1.

4. Majority-Rule Crowdvoting

In this section, we consider the simplest decision rule, which sets $w_i = 1$ for all i, giving

$$R_u(v) = \begin{cases} 1, & \sum_{i=1}^{n} v_i \geq m_p \\ -1, & \sum_{i=1}^{n} v_i \leq -m_q \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

Table 1. Notation

n	Number of voters for an assessment
$T \in \{-1, 1\}$	Underlying truth of an assessment
ϱ	Probability that $T = 1$: $P(T = 1)$
$v_i \in \{-1, 1\}$	Voter i's vote
p_i	Voter i's conditional accuracy in determining $T = 1$: $P(v_i = 1 \mid T = 1)$
q_i	Voter i's conditional accuracy in determining $T = -1$: $P(v_i = -1 \mid T = -1)$
ρ_i	Voter i's unconditional accuracy in determining T: $P(v_i = T)$
$w_i \in \{1, \rho_i\}$	Voter i's weight in the voting aggregation
m_p	Consensus parameter for concluding $T = 1$
m_q	Consensus parameter for concluding $T = -1$

where the u subscript refers to the unit weights; in other words, we are considering "majority-rule crowdvoting." The consensus parameters m_p and m_q can be used to capture proportional supermajorities. For example, if the system requires at least a fraction $\kappa \in (0, 1)$ of the n voters to vote $v = 1$ in order for the system to conclude that $T = 1$, the rule needs $\sum_{i=1}^{n} v_i \geq \kappa n - (1 - \kappa)n$ to conclude $T = 1$, and we can set $m_p = (2\kappa - 1)n$.

The probability that the crowd correctly determines the underlying truth of a given assessment,

$$P(R(v) = T), \tag{6}$$

is of critical importance in a crowdvoting system and is the primary service metric that this paper analyzes. We call this the PCCA. We consider two approaches for implementing the majority-rule crowdvoting rule in Equation (5), resulting in two expressions for the PCCA. In the first, we present a static analysis, in which the number of voters n is determined in advance of the random sampling of n voters. This analysis is analogous to sample size determination in statistical experiment design. We then consider a dynamic variant, using random walks, with which the number of voters n is not determined in advance. A main contribution of the results in this section (and the next section) is the derivation of *closed-form* expressions for the PCCA, which allow a manager to more easily understand the crowdvoting strategies, select among them, and more readily implement them. We then provide detailed analyses, based on the optimization of problem (2), to guide a user in appropriately setting the consensus parameters m_p and m_q for both approaches.

4.1. A Priori Determination of Number of Voters n

In this section, we assume that the number of voters n is fixed in advance of an assessment's evaluation by the crowd. Our first result is the following proposition.

Proposition 1. *If the p_i and q_i are i.i.d., the PCCA is*

$$P(R_u(v) = T) = \varrho\left(1 - F_{\mu_p}\left(\left\lceil\frac{n + m_p}{2}\right\rceil - 1\right)\right) + (1 - \varrho)\left(1 - F_{\mu_q}\left(\left\lceil\frac{n + m_q}{2}\right\rceil - 1\right)\right),$$

where F_μ is the cumulative distribution function of a binomial random variable with n trials and probability of success $\mu \in [0, 1]$.

Note that only the means μ_p and μ_q of the distributions f_p and f_q, respectively, are needed to evaluate the PCCA, which facilitates the implementation of this simple decision rule in practice. For example, the Matlab functions binocdf and ceil suffice as well as the Microsoft Excel functions binom.dist and roundup. In Appendix C, we demonstrate some counterintuitive nonmonotonicity results that occur if $\mu_p < 0.5$ or $\mu_q < 0.5$. Conveniently, this unsatisfactory behavior of the PCCA associated with $\mu_p < 0.5$ or $\mu_q < 0.5$ can be rather easily rectified. For instance, for any voter i who has $p_i < 0.5$, the system can invert the vote and improve the voter's accuracy; for this voter, the new vote v_i' is defined as $v_i' = -v_i$, which results in a modified accuracy $p_i' = \max\{p_i, 1 - p_i\} \geq 0.5$. In this way, we can guarantee $\mu_p > 0.5$ and $\mu_q > 0.5$, and the system can obtain the desirable behavior that the PCCA is monotonically increasing in (odd) n. The distributions f_p and f_q aree updated as well to have a domain $[a, b] \subseteq [0.5, 1]$, which does not affect our subsequent analysis (as discussed in Section 3.1). In practice, managers should make the effort to apply this modification to any voters that (perhaps intentionally) perform worse than a fair coin flip. We also provide alternative techniques, based on type I and II errors, to avoid these undesired behaviors by appropriately setting the consensus parameters m_p and m_q; we discuss this approach, and its limitations, in the next section.

Example 1. Proposition 1 has some interesting implications for academic peer review. For instance, if there are $n = 2$ reviewers and $\mu_p = \mu_q = 0.90$ (which we feel could be common in academic peer review), then $P(R_u(v) = T) = 0.81$, which does not instill much confidence in the peer-review process. Furthermore, *a single reviewer* can attain $P(R_u(v) = T) = 0.90$, so the second reviewer is actually detrimental. The intuition for this counterintuitive result is that there is a possibility of deadlock, in which one reviewer votes for and the other against a submitted paper, which destroys the value of their accuracies. To conclude these thoughts, we observe that a third reviewer gives $P(R_u(v) = T) = 0.972$, which is much more satisfying.

We next consider correlated voters. We were unable to find closed-from expressions, as in Proposition 1, for the case of nonzero correlations r_p and r_q; we, therefore, study this case numerically. We assume that the accuracy distributions f_p and f_q are uniformly distributed on $[0.5, 1.0]$, a distributional assumption that results in the clearest differentiation between correlation values. For a given value of n, we generate correlated accuracies using the Gaussian copula method (Nelsen 2007) for $r_p = r_q \in \{0.00, 0.50, 0.99\}$,[1] and the PCCA is evaluated using Monte Carlo simulation with 10,000 trials. Setting $m_p = m_q = 1$, in Figure 2 we observe that the PCCA *decreases* as the common correlation coefficient is increased (we only plot odd n to avoid the local minima at even n because of ties). Intuitively, this can be interpreted as a lack of diversity in the crowd, which has been identified as a key factor in crowdsourcing (e.g., see Terwiesch and Xu 2008). Although not apparent in Figure 2, the PCCA converges to unity for each correlation coefficient; however, the convergence is clearly slower as the correlation coefficient increases.

4.1.1. Type I and II Error Probabilities.

We next apply the idea of type I and II errors to our crowdvoting context so that we may control the probabilities of these errors. This analysis provides us a means to appropriately determine the consensus parameters m_p and m_q. We define a null hypothesis as an assessment that is not offensive: $T = 1$. The probabilities of type I and II errors, for i.i.d. accuracies, are characterized in the following proposition.

Proposition 2. *If the p_i and q_i are i.i.d., the probability of a type I error is*

$$P(R_u(v) = -1 | T = 1) = F_{\mu_p}\left(\left\lfloor \frac{n - m_q}{2} \right\rfloor\right),$$

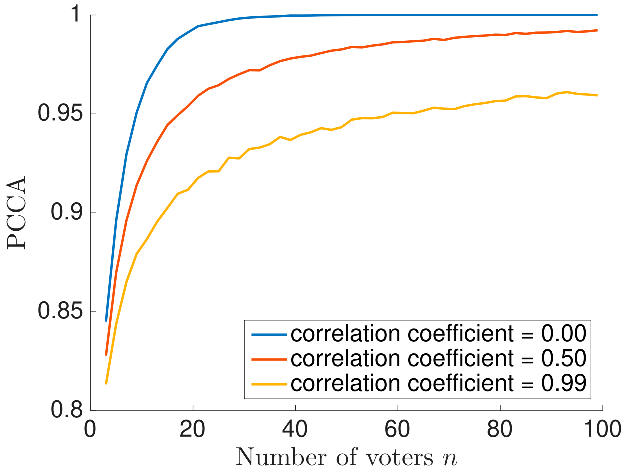
and the probability of a type II error is

$$P(R_u(v) = 1 | T = -1) = F_{\mu_q}\left(\left\lfloor \frac{n - m_p}{2} \right\rfloor\right),$$

where F_μ is the cumulative distribution function of a binomial random variable with n trials and probability of success $\mu \in [0, 1]$.

Suppose that we require the probability of a type I error to be at most $\epsilon \in (0, 0.5)$, the probability of a type II error to be at most $\varepsilon \in (0, 0.5)$, and a PCCA of at least $\gamma \in (0.5, 1)$. Note that Proposition 2 can be used to determine the values of m_p and m_q as a function of n to obtain these target error probabilities. In particular, if

Figure 2. The PCCAs for Various Correlation Coefficients



$k_n^* = \max\{k \in \mathbb{Z} | F_{\mu_p}(k) \le \epsilon\}$, then we set $m_q = n - 2k_n^*$. Similarly, if $\ell_n^* = \max\{\ell \in \mathbb{Z} | F_{\mu_q}(\ell) \le \varepsilon\}$, then we set $m_p = n - 2\ell_n^*$. We can then plug these consensus parameters into the PCCA expression from Proposition 1, obtaining $P(R_u(v) = T) = \varrho(1 - F_{\mu_p}(n - \ell_n^* - 1)) + (1 - \varrho)(1 - F_{\mu_q}(n - k_n^* - 1))$, which is increasing in n. We formalize these ideas in the following corollary.

Corollary 1. *If $m_q = n - 2k_n^*$ and $m_p = n - 2\ell_n^*$, where $k_n^* = \max\{k \in \mathbb{Z} | F_{\mu_p}(k) \le \epsilon\}$ and $\ell_n^* = \max\{\ell \in \mathbb{Z} | F_{\mu_q}(\ell) \le \varepsilon\}$, then*

$$P(R_u(v) = -1 | T = 1) \le \epsilon, \qquad P(R_u(v) = 1 | T = -1) \le \varepsilon,$$

and

$$P(R_u(v) = T) = \varrho\Big(1 - F_{\mu_p}\big(n - \ell_n^* - 1\big)\Big) + \big(1 - \varrho\big)\Big(1 - F_{\mu_q}\big(n - k_n^* - 1\big)\Big). \tag{7}$$

Therefore, we can (numerically) find the minimum number of voters needed to obtain a target PCCA $\ge \gamma$ while satisfying constraints on the probabilities of type I and II errors. In other words, we solve the following problem, which is a detailed version of problem (2):

$$\begin{aligned} \min_{n, m_p, m_q} \quad & n \\ \text{s.t.} \quad & P(R_u(v) = T) \ge \gamma \\ & P(R_u(v) = -1 | T = 1) \le \epsilon \\ & P(R_u(v) = 1 | T = -1) \le \varepsilon. \end{aligned} \tag{8}$$

Example 2. We demonstrate the procedure to solve this optimization problem using the parameter values $\epsilon = 0.01$, $\varepsilon = 0.01$, $\gamma = 0.98$, and $(\mu_p, \mu_q) = (0.8, 0.6)$: we obtain $(n, m_p, m_q) = (28, 8, -4)$. Note that *negative* consensus parameters (e.g., $m_q < 0$) are needed because of the relatively lower average accuracy μ_q in determining $T = -1$ as compared with μ_p in determining $T = 1$. Mathematically,

$$R_u(v) = \begin{cases} 1, & \sum_{i=1}^{28} v_i \ge 8 \\ -1, & \sum_{i=1}^{28} v_i \le 4 \\ 0, & \text{otherwise,} \end{cases}$$

and the rule concludes that $T = -1$ even if $\sum_{i=1}^{28} v_i \in \{0, 1, 2, 3, 4\}$! This rule gives $P(R_u(v) = T) = 0.9818$, $P(R_u(v) = -1 | T = 1) = 0.005$, $P(R_u(v) = 1 | T = -1) = 0.0081$. In contrast, if $n = 28$ and $(m_p, m_q) = (1, 1)$, then $P(R_u(v) = T) = 0.9064$, $P(R_u(v) = -1 | T = 1) \approx 0$, and $P(R_u(v) = 1 | T = -1) = 0.1015$; even if $n = 75$ and $(m_p, m_q) = (1, 1)$, then $P(R_u(v) = T) = 0.9802$, $P(R_u(v) = -1 | T = 1) \approx 0$, but $P(R_u(v) = 1 | T = -1) = 0.0396$.

We next consider correlated voters. Here, we introduce asymmetry by letting f_p be uniformly distributed on $[0.75, 1.0]$ and f_q be uniformly distributed on $[0.5, 1.0]$; we obtained qualitatively similar results with other distributional assumptions. We solve problem (8) numerically: we consider a grid in (m_p, m_q) space, and for each feasible tuple (satisfying the second and third bounds in (8)), we find the minimum value of n that satisfies the first bound in (8); finally, we find the smallest feasible n over all feasible (m_p, m_q) tuples. To evaluate feasibility, we again utilize the Gaussian copula method to generate correlated accuracies and Monte Carlo simulation to evaluate the constraint probabilities in problem (8). Table 2 reports on the optimal values of n, m_p, and m_q for $\epsilon = \varepsilon = 0.01$, $\gamma = 0.98$, $r_p = r_q \in \{0.00, 0.50, 0.99\}$, and $\varrho = 0.5$.

From Table 2, we again see that the asymmetry in distributions leads to a negative consensus parameter $m_q < 0$ at optimality, which is due to the relative crowd strength in identifying $T = 1$ over $T = -1$. Furthermore, the optimal (n, m_p, m_q) values preserve the property that stronger correlations weaken the crowd's ability to

Table 2. Optimal Solutions to Problem (8) for $\epsilon = \varepsilon = 0.01$ and $\gamma = 0.98$ and $r_p = r_q \in \{0.00, 0.50, 0.99\}$

(μ_p, μ_q)	$r_p = r_q = 0.0$	$r_p = r_q = 0.5$	$r_p = r_q = 0.99$
n	11	17	23
m_p	2	4	6
m_q	−2	−4	−3

correctly judge an assessment, thus leading to a higher number of voters per assessment for target values of PCCA and error probabilities.

4.2. Dynamic Determination of Number of Voters n via Random Walks

In this section, we assume that the number of voters n is not set in advance of an assessment's evaluation by the crowd. In other words, voters are added to an assessment until consensus is reached. We propose that *two* simple biased random walks can collectively serve as a dynamic implementation of the unit-weighted crowdvoting rule in Equation (5). To explain this link, note that the proof of Proposition 1 shows that, conditional on $T = 1$, voter i's vote v_i is a Rademacher random variable with parameter μ_p: $P(v_i = 1) = \mu_p$ and $P(v_i = -1) = 1 - \mu_p$. Similarly, conditional on $T = -1$, we have $P(v_i = 1) = 1 - \mu_q$ and $P(v_i = -1) = \mu_q$. The aggregation of the first k votes in Equation (5), $\sum_{i=1}^{k} v_i$, returns the position of a simple random walk after k steps, whose parameter depends on the value of T. In addition, assuming m_p and m_q are positive integers, we may define the stopping time as the number of voters n:

$$n = \min\left\{k \geq 1 : \sum_{i=1}^{k} v_i = m_p \text{ or } \sum_{i=1}^{k} v_i = -m_q\right\}. \tag{9}$$

Note that the application of a random walk effectively requires m_p and m_q to be positive (to avoid a trivial stopping time of zero), which is not required in the static rule studied in the previous section. Thus, this new rule suggests a trade-off between the benefit of dynamism and the restriction of consensus parameters, which we explore in this section.

Standard probability theory (e.g., Steele 2000, chapter 1) states that $P(n < \infty) = 1$ under independent voters, which implies that $P(R_u(v) = 0) = 0$. The PCCA under this dynamic implementation of Equation (5) is given in the following proposition.

Proposition 3. *If the p_i and q_i are i.i.d., the PCCA is*

$$P(R_u(v) = T) = \varrho\left(\frac{1 - \left(\frac{1-\mu_p}{\mu_p}\right)^{m_q}}{1 - \left(\frac{1-\mu_p}{\mu_p}\right)^{m_p+m_q}}\right) + (1-\varrho)\left(\frac{1 - \left(\frac{1-\mu_q}{\mu_q}\right)^{m_p}}{1 - \left(\frac{1-\mu_q}{\mu_q}\right)^{m_p+m_q}}\right).$$

Note that Proposition 3 does not depend on n, which varies across assessments. The *expected* number of voters is characterized in the next proposition.

Proposition 4. *If the p_i and q_i are i.i.d., the expected number of voters is*

$$E[n] = \varrho\left(\frac{m_q}{1-2\mu_p} - \left(\frac{m_p + m_q}{1-2\mu_p}\right)\left(\frac{1 - \left(\frac{1-\mu_p}{\mu_p}\right)^{m_q}}{1 - \left(\frac{1-\mu_p}{\mu_p}\right)^{m_p+m_q}}\right)\right)$$
$$+ (1-\varrho)\left(\frac{m_p}{1-2\mu_q} - \left(\frac{m_p + m_q}{1-2\mu_q}\right)\left(\frac{1 - \left(\frac{1-\mu_q}{\mu_q}\right)^{m_p}}{1 - \left(\frac{1-\mu_q}{\mu_q}\right)^{m_p+m_q}}\right)\right).$$

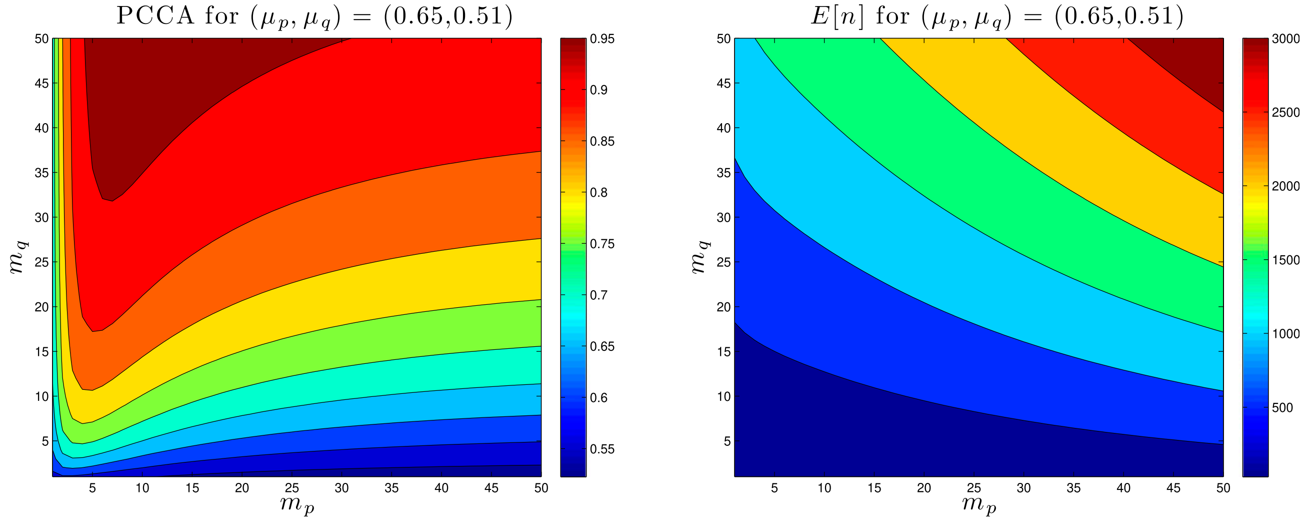
Under this random walk model, we analyze the following variant of problem (2), in which we determine m_p and m_q to minimize the expected number of voters, subject to a service-level constraint γ on the PCCA:

$$\begin{aligned} \min_{m_p, m_q} \quad & E[n] \\ \text{s.t.} \quad & P(R_u(v) = T) \geq \gamma \\ & m_p, m_q \geq 0; \end{aligned} \tag{10}$$

note that the type I and II errors are simply $1 - \gamma$ because, as we discussed, $P(R_u(v) = 0) = 0$.

We found the derivation of a closed-form solution to problem (10) intractable. Fortunately, a numerical solution is straightforward to implement. In Figure 3, we display the contours of the feasible region on the left panel and the contours of the objective function on the right panel; we selected the means $(\mu_p, \mu_q) = (0.65, 0.51)$ to demonstrate that the feasible region is not necessarily convex, which supports our numerical solution procedure.

Figure 3. The PCCA (Left) and $E[n]$ (Right) as a Function of m_p and m_q Under the Random Walk Model



We next compare the dynamic determination of n considered in this section with the predetermination of n studied in the previous section under the assumption of independent voters. First, to attain any target PCCA value of $\gamma \in (0, 1)$, the random walk requires $\mu_p > 0.5$ and $\mu_q > 0.5$ because $m_p, m_q > 0$; fortunately, this is easily achievable by inverting the votes of inaccurate voters as described in the previous section. Consequently, we consider the values $(\mu_p, \mu_q) = (0.8, 0.6)$ (the distributions are not needed). From Example 1, predetermination of n resulted in $n = 28$ voters, $m_p = 8$ and $m_q = -4$. The solution of problem (10) for $(\mu_p, \mu_q) = (0.8, 0.6)$ is $m_p = 9$, $m_q = 3$, and the minimized $E[n] = 54$. Thus, on average, the random walk implementation requires almost twice as many voters. Table 3 provides a more comprehensive comparison that shows predetermination of n, in all cases, results in fewer voters, on average, than the dynamic determination of n. The difference in the number of voters increases with $|\mu_p - \mu_q|$. The reason for this disadvantage is that the random walk implementation requires the consensus parameters m_p and m_q to be positive, whereas the predetermination of n does not; indeed, as mentioned, $m_q = -4$ for the predetermination of n when $(\mu_p, \mu_q) = (0.8, 0.6)$, which allows the crowdvoting rule in Equation (5) to better accommodate differential accuracies; the random walk implementation is limited in its ability to do so by the constraints $m_p, m_q \geq 0$. Thus, although one might think the dynamic implementation is more efficient than the static one, we recommend that managers use the static predetermination of n because of the advantage of negative consensus parameters. Similar results were obtained under correlated voters (i.e., predetermination is better) and are omitted for brevity.

4.3. A Hybrid Voting Mechanism

In this section, we provide a hybrid voting mechanism that combines the dynamic structure of the mechanism analyzed in Section 4.2 with the more flexible consensus parameters of the static mechanism studied in Section 4.1.

Beginning with target PCCA and type I and II error probabilities of $(\gamma, \epsilon, \varepsilon)$, the analysis in Section 4.1 prescribes the appropriate values of (n^s, m_p^s, m_q^s), where the s superscript refers to the static setting. However, the same consensus conclusions can potentially be achieved with fewer voters in a dynamic setting even with negative consensus parameters. For instance, if the first $n^s - l$ voters achieve a consensus of $m_p^s + l$, we readily conclude that the consensus of all n^s voters must be at least m_p^s; similarly, if the first $n^s - l$ voters achieve a consensus of $-m_q^s - l$, the consensus of all n^s voters is at most $-m_q^s$. We can exploit this

Table 3. Minimum Number of Voters to Obtain $P(R(v) = T) \geq 0.98$ for Various (μ_p, μ_q)

(μ_p, μ_q)	$(0.6, 0.6)$	$(0.7, 0.6)$	$(0.8, 0.6)$	$(0.9, 0.6)$	$(0.9, 0.7)$	$(0.9, 0.8)$	$(0.9, 0.9)$
n (Section 4.1)	122	54	28	16	12	9	5
$E[n]$ (Section 4.2)	146	74	54	40	18	11	7
$E[n]$ (Section 4.3)	103	43	21	12	8	6	3

observation in a dynamic setting: modifying Equation (9), a hybrid voting mechanism dynamically stops at n voters, where

$$n = n^s - \max\left\{ l \in \{0, \dots, n^s - 1\} : \sum_{i=1}^{n^s - l} v_i = m_p^s + l \text{ or} \sum_{i=1}^{n^s - l} v_i = -m_q^s - l \right\},$$

where the last l votes are not needed. In Table 3, we report the expected number of voters of this hybrid strategy; comparing these results with those of Sections 4.1 and 4.2, we see that this hybrid strategy exhibits the strongest performance by combining dynamism and negative consensus parameters.

5. Accuracy-Weighted Crowdvoting

In this section, we analyze a decision rule in which a vote's weight is equal to the voter's accuracy. The motivation is to give more accurate voters a stronger influence in determining the system's assessment. Setting $w_i = \rho_i$, where ρ_i is given in Equation (3), the decision rule in this section is

$$R_\rho(v) = \begin{cases} 1, & \sum_{i=1}^{n} \rho_i v_i \geq m_p \\ -1, & \sum_{i=1}^{n} \rho_i v_i \leq -m_q \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

where the ρ subscript indicates the accuracy weights $w_i = \rho_i$. As discussed previously, the weight ρ_i is random though easily implemented in practice with known values of p_i, q_i, and ϱ.

We again consider two approaches, one static and the other dynamic, for implementing the accuracy-weighted crowdvoting rule in Equation (11), resulting in different expressions for the PCCA. As in Section 4, the first considers the predetermination of n, and the second allows n to be determined dynamically. We provide analytical solutions under the assumption of independent voters, which we extend via numerical studies for correlated voters.

5.1. A Priori Determination of Number of Voters *n*

Our first step in understanding this new decision rule is to analyze the random variable $\rho_i v_i$. Note that this random variable is continuous, which implies that the probability of a tie vote is zero, so there is no need to avoid an even number of voters (as in simple majority-rule voting). The next lemma characterizes the conditional distributions of this random variable.

Lemma 1. *The density of $\rho_i v_i$, conditional on $T = 1$, is*

$$g(t) = \begin{cases} t f_p(t), & t \in [0, 1] \\ (1 + t) f_p(-t), & t \in [-1, 0), \end{cases}$$

and the density, conditional on $T = -1$, is

$$h(t) = \begin{cases} (1 - t) f_q(t), & t \in [0, 1] \\ -t f_q(-t), & t \in [-1, 0), \end{cases}$$

for $i = 1, \dots, n$.

In theory, Lemma 1 suffices to evaluate the PCCA for given values of n, m_p, and m_q under independent voters. The density $\hat{g}_n$ of $\sum_{i=1}^{n} \rho_i v_i$, conditional on $T = 1$, can be determined by the n-fold convolution $(g * g * \cdots * g)(t)$, where g is the conditional density of $\rho_i v_i$ from Lemma 1, and convolution is defined as $(g * g)(t) = \int_{-\infty}^{\infty} g(\tau) g(t - \tau) d\tau$. Similarly, the density $\hat{h}_n$ of $\sum_{i=1}^{n} \rho_i v_i$, conditional on $T = -1$, can be determined by the n-fold convolution of the density h from Lemma 1. The PCCA can then be evaluated as

$$\begin{aligned} P\big(R_\rho(v) = T\big) &= \varrho P\big(R_\rho(v) = 1 | T = 1\big) + \big(1 - \varrho\big) P\big(R_\rho(v) = -1 | T = -1\big) \\ &= \varrho \int_{m_p}^{n} \hat{g}_n(\tau) d\tau + \big(1 - \varrho\big) \int_{-n}^{-m_q} \hat{h}_n(\tau) d\tau. \end{aligned} \tag{12}$$

We can utilize numerical convolution to determine the conditional densities $\hat{g}_n$ and $\hat{h}_n$ for $\sum_{i=1}^n \rho_i v_i$ and numerical integration to evaluate $P(R_\rho(v) = T)$, which we illustrate in the following example.

Example 3. Again consider the context of academic peer review with $n = 2$ reviewers and suppose that f_p and f_q are both uniform distributions on $[0.8, 1]$, where $\mu_p = \mu_q = 0.9$. From Example 1, the value of $P(R_u(v) = T) = 0.81$ and the value of $P(R_p(v) = T) = 0.93$. Therefore, moving from a unit-weighted voting rule to an accuracy-weighted rule, by incorporating reviewer skills (estimable from previous peer reviews), can substantially increase $P(R(v) = T)$ by 15%.

However, this numerical approach is cumbersome and susceptible to numerical problems, which limits the applicability and adoption of this research in practice. Therefore, we also provide approximations and characterize their quality (in Appendix D). In particular, we use the central limit theorem (CLT) to approximate $\sum_{i=1}^n \rho_i v_i$ for both $T = 1$ and $T = -1$, and obtain the following result.

Proposition 5. *If the p_i and q_i are i.i.d.,*

$$P(R_\rho(v) = T) \approx \varrho\Phi\left(\frac{n\left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right) - m_p}{\sqrt{n}\sqrt{\sigma_p^2 + \mu_p^2 - \left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)^2}}\right) + (1 - \varrho)\Phi\left(\frac{n\left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right) - m_q}{\sqrt{n}\sqrt{\sigma_q^2 + \mu_q^2 - \left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right)^2}}\right),$$

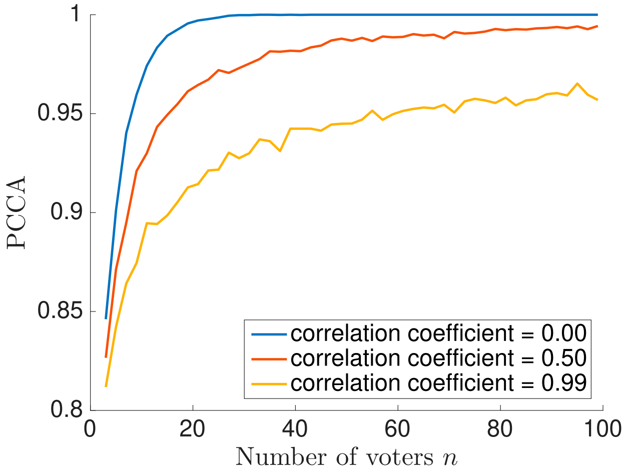
where Φ is the standard normal distribution. Furthermore, if $2(\sigma_p^2 + \mu_p^2) > \mu_p$ and $2(\sigma_q^2 + \mu_q^2) > \mu_q$, then $\lim_{n\to\infty} P \times (R_\rho(v) = T) = 1$.

The approximation for the PCCA for finite n can be easily implemented in Matlab (the `normcdf` function) or Microsoft Excel (the `norm.s.dist` function). We also point out that, in contrast to Proposition 1, which only requires the means μ_p and μ_q of the distributions f_p and f_q, respectively, the approximation in Proposition 5 also requires the standard deviations σ_p and σ_q. However, this latter information requirement is still less demanding than that for calculating the exact probability, via numerical convolution and integration, which requires the full distributions f_p and f_q.

As in Section 4.1, we explore the accuracy-weighted voting rule under correlated voters. Using the same methodology previously explained except applied to Equation (11), we obtain very similar results to that for the majority-rule mechanism, which is presented in Figure 4. In particular, for a given number of voters, the PCCA is reduced as the correlation coefficient is increased. However, there are differences between Figures 4 and 2 in that the former is effectively a shifted (upward) version of the latter; in other words, for a given correlation coefficient and number of voters n, the PCCA for the accuracy-weighted rule is larger than that for majority rule. We explore the precise benefit of accuracy votes later in this section via an appropriate version of problem (2).

5.1.1. Type I and II Error Probabilities. We continue our analysis by determining approximations for the probabilities of type I and II errors.

Figure 4. The PCCAs for Various Correlation Coefficients



Proposition 6. *If the p_i and q_i are i.i.d., the probability of a type I error is*

$$P\left(R_\rho(v) = -1 | T = 1\right) \approx \Phi\left(\frac{-m_q - n\left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)}{\sqrt{n}\sqrt{\sigma_p^2 + \mu_p^2 - \left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)^2}}\right)$$

and the probability of a type II error is

$$P\left(R_\rho(v) = 1 | T = -1\right) \approx \Phi\left(\frac{-m_p - n\left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right)}{\sqrt{n}\sqrt{\sigma_q^2 + \mu_q^2 - \left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right)^2}}\right),$$

where Φ is the standard normal distribution.

As in Section 4.1, we demonstrate how to determine the minimum number of voters n to obtain any desired PCCA while limiting the probabilities of type I and II errors. Again suppose that we require the probability of a type I error to be at most $\epsilon \in (0, 0.5)$, the probability of a type II error to be at most $\varepsilon \in (0, 0.5)$, and a PCCA of at least $\gamma \in (0.5, 1)$. We, thus, solve the following variant of problem (2):

$$\begin{aligned} \min_{n, m_p, m_q} \quad & n \\ \text{s.t.} \quad & P\left(R_\rho(v) = T\right) \geq \gamma \\ & P\left(R_\rho(v) = -1 | T = 1\right) \leq \epsilon \\ & P\left(R_\rho(v) = 1 | T = -1\right) \leq \varepsilon. \end{aligned} \tag{13}$$

Proposition 6 can be used to determine values of m_p and m_q as a function of n to attain these target error probabilities. Observing that

$$\Phi\left(\frac{-m_q - n\left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)}{\sqrt{n}\sqrt{\sigma_p^2 + \mu_p^2 - \left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)^2}}\right) \leq \epsilon$$

$$\Leftrightarrow \quad -m_q \leq n\left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right) + \sqrt{n}\sqrt{\sigma_p^2 + \mu_p^2 - \left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)^2}\,\Phi^{-1}(\epsilon),$$

we set

$$m_q = -n\left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right) - \sqrt{n}\sqrt{\sigma_p^2 + \mu_p^2 - \left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)^2}\,\Phi^{-1}(\epsilon).$$

Similarly, we set

$$m_p = -n\left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right) - \sqrt{n}\sqrt{\sigma_q^2 + \mu_q^2 - \left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right)^2}\,\Phi^{-1}(\varepsilon).$$

Plugging these consensus parameters into the PCCA expression from Proposition 5, we are able to determine the minimum number of voters needed to meet our target probabilities.

Example 4. We demonstrate this procedure using the parameter values $\epsilon = \varepsilon = 0.01$ and $\gamma = 0.98$. We let f_p and f_q be beta distributions with $\beta = 1$, and we choose α's so that the means of the beta distributions match those of Example 1, namely $(\mu_p, \mu_q) = (0.8, 0.6)$. Applying the preceding procedure, we obtain the optimal solution to (13) as $(n, m_p, m_q) = (12, 1.7659, -1.4178)$; recall that, in Example 1, the majority-rule mechanism required $(n, m_p, m_q) = (28, 8, -4)$, and we conclude that fewer voters are required for the accuracy-weighted rule for a given $(\gamma, \epsilon, \varepsilon)$ combination. To complete this example, we also calculate the *exact* PCCA for the accuracy-weighted rule for our optimal parameters, calculated using numerical convolution and integration with a discretization step size of $\delta = 0.0001$ and we determine that the PCCA $= 0.9787 \approx \gamma$, providing evidence that our approximations are of high quality (further evaluations of the approximations can be found in Appendix D).

We now consider correlated voters via numerical studies as in Section 4.1. We again assume that f_p is uniformly distributed on $[0.75, 1.0]$ and f_q is uniformly distributed on $[0.5, 1.0]$. We solve problem (13) numerically using a similar approach to that outlined for the solution of problem (8). Table 4 reports on the

Table 4. Optimal Solutions to Problem (8) for $\epsilon = \varepsilon = 0.01$ and $\gamma = 0.98$ and $r_p = r_q \in \{0.00, 0.50, 0.99\}$

(μ_p, μ_q)	$r_p = r_q = 0.0$	$r_p = r_q = 0.5$	$r_p = r_q = 0.99$
n	9 (11)	13 (17)	17 (23)
m_p	0.90 (2)	1.90 (4)	2.90 (6)
m_q	−0.80 (−2)	−1.10 (−4)	−2.00 (−3)

optimal values of n, m_p, and m_q for $\epsilon = \varepsilon = 0.01$, $\gamma = 0.98$, $r_p = r_q \in \{0.00, 0.50, 0.99\}$, and $\varrho = 0.5$. To facilitate comparisons between the solutions to problems (8) and (13), we also list, in parentheses, the optimal values of n, m_p, and m_q from Table 2.

Table 4 shows that the accuracy-weighted voting rule requires fewer voters than the unit-weighted voting rule to attain a target PCCA for given limits on the error probabilities. Furthermore, the gap increases with the common correlation coefficient, and we may conclude that the accuracy-weighted rule is more adept at handling correlated voters. However, there is a clear trade-off as the accuracy-weighted rule is more difficult to implement. An exact implementation, which fully leverages the distributions f_p and f_q, requires sophisticated computational analysis that is susceptible to numerical problems. An approximate implementation that only requires the means and standard deviations of the distributions f_p and f_q is potentially inaccurate; however, the quality of the approximations in Table 4 is encouraging. These approaches, centered around exact and approximate expressions for the PCCA, are only applicable for independent voters; correlated voters require a fully numerical solution approach that combines generation of correlated accuracies via the Gaussian copula method, Monte Carlo simulations to evaluate the appropriate probabilities, and numerical optimization over a grid of variable values. Finally, the estimation of the accuracies p_i and q_i for each voter i can be burdensome as discussed in Section 3.1. Thus, although the performance of the accuracy-weighted rule is clearly superior, the implementation difficulties are not negligible, and a manager might settle for the simpler majority-rule voting mechanism because of ease of implementation.

5.2. Dynamic Determination of Number of Voters *n* via Sequential Hypothesis Testing

In this section, we consider an analogue to the dynamic determination of the number of voters n via random walks in Section 4.2 for unit weights under independent voters. Here, for accuracy weights, the aggregation of the first k votes in Equation (11), $\sum_{i=1}^{k} x_i$, where $x_i = \rho_i v_i$, returns the position of a random walk after k steps. Note that the walk is no longer *simple* because the x_i are continuously distributed on $[-1, 1]$ rather than on $\{-1, 1\}$, according to one of the distributions, g or h, in Lemma 1. We may define the stopping time as the number of voters n, where one of the consensus parameters is *exceeded*,

$$n = \min\left\{k \geq 1 : \sum_{i=1}^{k} x_i \geq m_p \text{ or } \sum_{i=1}^{k} x_i \leq -m_q\right\}, \tag{14}$$

where m_p and m_q are positive values (again, to preclude a trivial stopping time) though not necessarily integer. Unfortunately, to the best of our knowledge, analogues of the closed-form expressions in Propositions 3 and 4 do not exist for random walks with continuously distributed step sizes $x_i = \rho_i v_i$. Therefore, we take a different, yet related, approach based on the concept of *sequential hypothesis testing* as pioneered by Wald (1945). The core idea of this approach is to ascertain whether distribution g or h from Lemma 1 generates the weighted votes $x_i = \rho_i v_i$. Therefore, the technique works best when g and h are dissimilar. Also note that, even if f_p and f_q are identical, g and h are different.

We formally define our hypotheses as $H_0 : T = 1$ and $H_1 : T = -1$, which coincides with our previous definitions of type I and II errors; note that, under H_0, Lemma 1 indicates that $x_i = \rho_i v_i$ has density g, and under H_1, the density is h. Slightly abusing notation (for clarity), assume that k data points, x_i, $i = 1, \ldots, k$, are available. Standard (static) hypothesis testing defines the likelihood ratio as

$$\Lambda(x_1, \ldots, x_k) = \prod_{i=1}^{k} \frac{h(x_i)}{g(x_i)}. \tag{15}$$

This ratio is compared with a threshold η; if $\Lambda(x_1, \ldots, x_k) \geq \eta$, H_0 is rejected; otherwise, H_0 is accepted.

In *sequential* hypothesis testing, a third option of generating more data is available. In particular, there exist two thresholds $0 < B < A < \infty$ so that hypothesis testing can be implemented in a dynamic manner:

$$\begin{cases} \text{if } \Lambda(x_1, \ldots, x_k) \leq B, & \text{accept } H_0 \\ \text{if } \Lambda(x_1, \ldots, x_k) \geq A, & \text{accept } H_1 \\ \text{if } B < \Lambda(x_1, \ldots, x_k) < A, & \text{generate sample } x_{k+1} \text{ and repeat.} \end{cases}$$

By calculating the log-likelihood ratio

$$\log(\Lambda(x_1, \ldots, x_k)) = \sum_{i=1}^{k} \log\left(\frac{h(x_i)}{g(x_i)}\right),$$

and letting $z_i = \log(\frac{h(x_i)}{g(x_i)})$, we may interpret $\sum_{i=1}^{k} z_i$ as a random walk. We introduce an alternative definition for the stopping time n (i.e., number of voters) in terms of this new random walk:

$$n = \min\left\{k \geq 1 : \sum_{i=1}^{k} z_i \geq \log(A) \text{ or } \sum_{i=1}^{k} z_i \leq \log(B)\right\}. \tag{16}$$

Wald (1945) showed that, as long as the x_i are independent, $P(n < \infty) = 1$, which implies that $P(R_\rho(v) = 0) = 0$; we referenced a similar result for simple random walks in Section 4.2.

The classic analysis of sequential hypothesis testing is intimately linked with target probabilities of type I and II errors, which is convenient for our analysis. In particular, requiring $P(R_\rho(v) = -1 | T = 1) \leq \epsilon$ and $P(R_\rho(v) = 1 | T = -1) \leq \varepsilon$, Wald (1945) suggests using the boundaries

$$A = \frac{1 - \varepsilon}{\epsilon} \qquad \text{and} \qquad B = \frac{\varepsilon}{1 - \epsilon}, \tag{17}$$

which have come to be known as the Wald boundaries (Ghosh and Sen 1991). Note that there is a subtle approximation in the derivation of these boundaries, which is why we can sidestep the lack of relevant analogues to Propositions 3 and 4 for a random walk with continuously distributed steps x_i; fortunately, it has been argued (Wald 1945, Ghosh and Sen 1991) that the approximation is high quality as long as the expected value of z_i is small with respect to $\log(B)$ and $\log(A)$. Indeed, the decision rule associated with (16) combined with the Wald boundaries (17) is known as the *sequential probability ratio test* and has been proved under quite general settings to be optimal among all sequential tests in the sense that the expected stopping time is minimal for target probabilities of type I and II errors; see Wald and Wolfowitz (1948) for further details. The PCCA and the expected number of voters are given in the next two propositions.

Proposition 7. *The PCCA is*

$$P\big(R_\rho(v) = T\big) \approx \varrho(1 - \epsilon) + \big(1 - \varrho\big)(1 - \varepsilon).$$

Proposition 8. *The expected number of voters is*

$$E[n] \approx \varrho\left(\frac{\epsilon \log\left(\frac{1-\varepsilon}{\epsilon}\right) + (1 - \epsilon)\log\left(\frac{\varepsilon}{1-\epsilon}\right)}{\int_{-1}^{1} \log\left(\frac{h(t)}{g(t)}\right) g(t)dt}\right) + (1 - \varrho)\left(\frac{(1 - \varepsilon)\log\left(\frac{1-\varepsilon}{\epsilon}\right) + \varepsilon \log\left(\frac{\varepsilon}{1-\epsilon}\right)}{\int_{-1}^{1} \log\left(\frac{h(t)}{g(t)}\right) h(t)dt}\right).$$

Table 5. Number of (Expected) Voters n to Obtain $\gamma = 0.98$ and $\epsilon = \varepsilon = 0.01$ for Various (μ_p, μ_q), Where f_p and f_q Are Beta Distributions

(μ_p, μ_q)	$(0.6, 0.6)$	$(0.7, 0.6)$	$(0.8, 0.6)$	$(0.9, 0.6)$	$(0.9, 0.7)$	$(0.9, 0.8)$	$(0.9, 0.9)$
$E[n]$ (Section 5.2)	5	5	5	5	4	3	2
n (Section 5.1)	27	19	12	7	6	4	3
$E[n]$ (Section 4.3)	103	43	21	12	8	6	3
$E[n]$ (Section 4.2)	146	74	54	40	18	11	7
n (Section 4.1)	122	54	28	16	12	9	5

Table 6. Number of (Expected) Voters n to Obtain $\gamma = 0.98$ and $\epsilon = \varepsilon = 0.01$ for Various (μ_p, μ_q), Where f_p Is a Uniform Distribution on $[\mu_p - 0.1, \mu_p + 0.1]$ and f_q Is a Uniform Distribution on $[\mu_q - 0.1, \mu_q + 0.1]$

(μ_p, μ_q)	(0.6, 0.6)	(0.7, 0.6)	(0.8, 0.6)	(0.9, 0.6)	(0.9, 0.7)	(0.9, 0.8)	(0.9, 0.9)
$E[n]$ (Section 5.2)	41	1	1	1	1	1	3
n (Section 5.1)	104	43	20	9	7	5	3

In the remainder of this section, we present the results of computational studies that compare the (expected) number of voters required for (a) static implementation of simple majority voting (Section 4.1), (b) dynamic implementation of simple majority voting (Section 4.2), (c) the hybrid strategy of Section 4.3, (d) static implementation of accuracy-weighted voting (Section 5.1), and (e) dynamic implementation of accuracy-weighted voting (Section 5.2). In Table 5, we consider the case in which f_p and f_q are beta distributions with various means. Similarly, in Table 6, we consider the case in which f_p and f_q are uniformly distributed with various means. Note that the results from Sections 4.1–4.3 are omitted from Table 6 because they are the same as those in Table 5. In other words, although the distributions are different, the means are the same, and hence, the number of voters is the same for simple majority-rule crowdvoting because these simpler strategies only require the means.

Tables 5 and 6 show us that the crowdvoting implemented via sequential hypothesis testing results in the fewest *expected* number of voters in all cases tested for independent voters. For high-accuracy voter populations (e.g., last column of Table 6), the static n is comparable to $E[n]$, so the static approach is perhaps preferred because it is a deterministic guarantee. In contrast, for medium- to low-accuracy populations, $E[n]$ is much smaller than n, so the sequential hypothesis testing approach is clearly preferred.

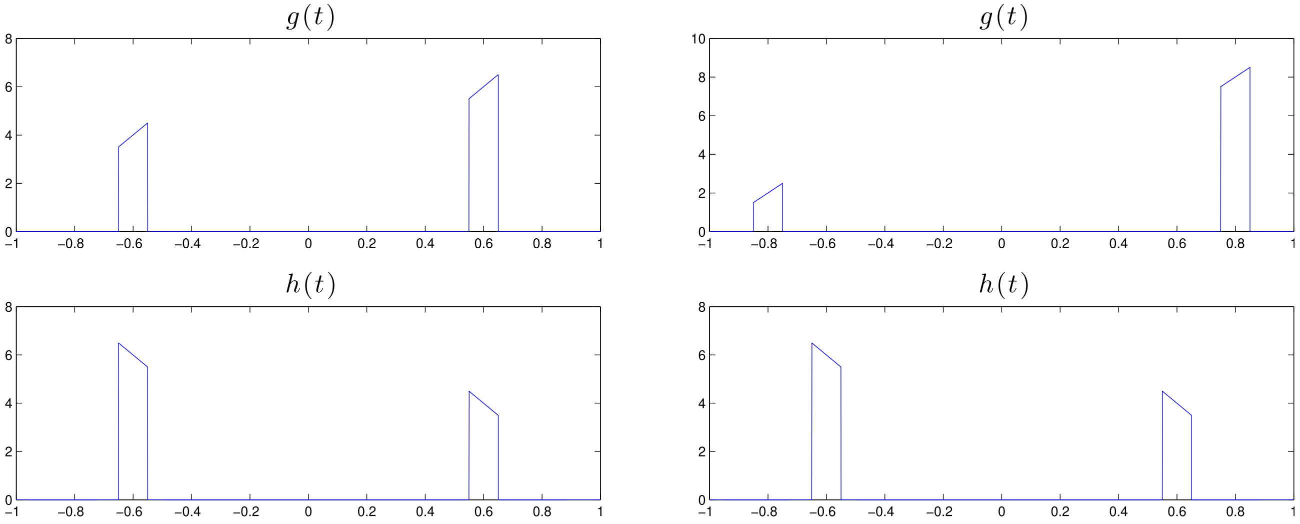
We also shed some light on the cases in which sequential hypothesis testing had difficulties; in the left panel of Figure 5, we present the densities g and h when f_p and f_q are uniformly distributed with means (0.6, 0.6), and it can be seen that the densities are similar, resulting in $E[n] = 41$; in the right panel, we present g and h when the means are (0.8, 0.6), which have mutually exclusive domains in which the densities are positive, resulting in $E[n] = 1$.

However, circumstances exist in which sequential hypothesis testing is not applicable, namely the case in which the densities g and h are equal. The simplest case that demonstrates this is possible is when the crowd simply flips a fair coin when voting; in other words, $p_i = q_i = 1/2$ deterministically, which we can represent using Dirac delta functions: $f_p(t) = \delta(t - 0.5)$ and $f_q(t) = \delta(t - 0.5)$. Applying Lemma 1, we obtain

$$g(t) = \begin{cases} t\delta(t-0.5), & t \in [0,1] \\ (1+t)\delta(-t-0.5), & t \in [-1,0) \end{cases} \quad \text{and} \quad h(t) = \begin{cases} (1-t)\delta(t-0.5), & t \in [0,1] \\ -t\delta(-t-0.5), & t \in [-1,0), \end{cases}$$

and we see that $g(t) = h(t)$ for all $t \in [-1, 1]$.

Figure 5. The Densities g and h When f_p and f_q are Uniformly Distributed with Means (0.6, 0.6) (Left Panel) and (0.8, 0.6) (Right Panel) Under the Sequential Hypothesis Testing Model



We also comment on correlated voters. Unfortunately, there is no existing literature (to the best of our knowledge) for applying sequential hypothesis testing to correlated data, and developing this theory is outside the scope of this paper. Therefore, we are unable to apply this dynamic technique to correlated voters, and we recommend using the numerical approach discussed in Section 5.1 when voters are strongly correlated.

Finally, we comment on practicality. Although the dynamic version of accuracy-weighted voting has the strongest performance of all voting mechanisms considered in this paper (with a few exceptions, detailed earlier), it is also undoubtedly the most difficult to implement in practice as it requires the g and h densities, which, in turn, require full knowledge of the f_p and f_q densities. The portfolio of voting mechanisms studied in this paper provides a manager a menu of options of increasing performance associated with increasing difficulty of implementation. Thus, a manager can adopt the most complex and highest performing voting mechanism that can be realistically implemented given the specific context and behavior of their online community members.

6. System Considerations and Limitations of Self Policing

In this section, we synthesize the results of all the previous sections to provide guidelines for designing a cost-effective crowdvoting system. We begin by assuming that an assessment must be resolved within a predetermined time from when the complaint was generated. For example, in the Xbox Live service, the desired turnaround time to provide a response to the complaint is 24 hours.

6.1. Supply and Demand of Assessments

In this section, we first elaborate upon volunteer voter participation, which provides a supply of (voters willing to evaluate) assessments. However, although a single voter might participate even without financial compensation, the voter's participation is clearly not unlimited, and therefore, the supply of voters is not unlimited; see Section 2.1 for further details of voter participation. For concreteness, suppose that M gamers are logged in and available to perform assessments at some point during a given 24-hour period. Note that M can be substantially less than the size of the online community (48 million in the Xbox Live case) because the gamers have to be logged in during a given 24-hour period. Assuming that each of the M gamers logs in once during the 24-hour period and is willing to evaluate m assessments on average, there is a supply of Mm votes available during these 24 hours; Section 2.1 outlines how a gamer would determine m.

Considering the demand for assessments, suppose that there are N assessments that need to be resolved during the 24-hour period. The techniques detailed in Sections 4.1–4.3, 5.1, and 5.2 allow us to derive the (expected) number of voters needed under each of the voting rules for target values of error probabilities and a target PCCA value. For ease of exposition, we let n represent the (expected) number of voters required for a single assessment under one of the techniques analyzed in this paper. This implies that we have a demand for Nn total votes in the 24-hour period.

A system cannot necessarily rely completely on the crowd to assess all content. For instance, if $Nn > Mm$, then there are not enough votes available to vote on all assessments, and the firm must rely on (costly) internal experts; note that $Nn > Mm$ is always possible if the target error probabilities are chosen to be small enough and the target PCCA is chosen to be high enough. Furthermore, if a crowd vote is inconclusive ($R(v) = 0$), then a firm expert is likely to be used. In other words, the employee assessment cost in problem (1) is positive at optimality. We let C denote the cost for an internal expert to evaluate one assessment; for example, assuming that an internal expert's salary is \$20/hour and each assessment takes 30 seconds, $C =$\$20/120 = \$0.17. The expected 24-hour firm cost (i.e., the employee assessment cost) for evaluating all assessments is, therefore,

$$cost_{24} = C\left(\left\lceil \frac{\max\{Nn - Mm, 0\}}{n} \right\rceil + \left\lfloor \frac{\min\{Mm, Nn\}}{n} \right\rfloor P(R(v) = 0)\right). \tag{18}$$

The expression $\max\{Nn - Mm, 0\}$ represents the vote shortfall, $\lceil \frac{\max\{Nn-Mm,0\}}{n} \rceil$ represents the corresponding number of assessments that an internal expert must evaluate, $\lfloor \frac{\min\{Mm,Nn\}}{n} \rfloor$ is the number of assessments assigned to the crowd, and $\lfloor \frac{\min\{Mm,Nn\}}{n} \rfloor P(R(v)=0)$ represents the expected number of inconclusive crowd assessments that must be handled by a firm expert. Note that the dynamic determinations of n considered in Sections 4.2 and 5.2 result in $P(R(v)=0)=0$.

6.2. Suggested Type I and II Error Probability Structure

Suppose that the probability of a type I error is required to equal $\epsilon \in (0,0.5)$, the probability of a type II error $\varepsilon \in (0,0.5)$, and a PCCA equal to $\gamma \in (0.5,1)$. Note that, if $\epsilon = \varepsilon = 1-\gamma$, then

$$\begin{aligned} P(R(v)=0) &= 1 - P\big(R(v)=T\big) - P\big(R(v)=-T\big) \\ &= 1-\gamma-\Big(\varrho P\big(R(v)=-1|T=1\big)+\big(1-\varrho\big)P\big(R(v)=1|T=-1\big)\Big) \\ &= 0. \end{aligned}$$

Therefore, we recommend the parameters $(\epsilon,\varepsilon,\gamma)$ be set in this way so that all crowd assessments are conclusive (not necessarily correct), and any assessment that goes to the crowd never goes to a costly firm expert. In addition, the three-parameter design vector $(\epsilon,\varepsilon,\gamma)$ simplifies to a scalar parameter design $(1-\gamma,1-\gamma,\gamma)$, a convenience for practice. Under these parameters, the 24-hour cost in Equation (18) simplifies to

$$cost_{24} = C\left\lceil \frac{\max\{Nn-Mm,0\}}{n} \right\rceil, \tag{19}$$

which is only positive if there are not enough crowd votes available. Of course, this cost can always be positive if γ is selected close enough to one.

6.3. Limits of Crowdvoting

In this section, we determine the limits of crowdvoting. In particular, we leverage the error probability structure of the previous section, $(\epsilon,\varepsilon,\gamma)=(1-\gamma,1-\gamma,\gamma)$, and the resulting cost expression in Equation (19) to determine the maximum value of the PCCA (γ) that can be attained *at zero cost*. In other words, we want to find the limit of a system in which no firm experts are utilized and the crowd assesses all complaints conclusively. If this value of γ is too low, Equation (19) can be used to determine the cost of a target PCCA.

For simplicity, we focus on the majority-vote rule with a static determination of n (Section 4.1); an analysis for the other voting rules is analogous. As an example, we consider average accuracies $(\mu_p,\mu_q)=(0.8,0.6)$ and set $\varrho=0.5$. We let $M=5{,}000{,}000$, approximately 10% of the total population on the Xbox Live service, and set $m=5$. We let $N=\nu M$, where $\nu \in \{50\%,25\%,10\%,1\%\}$, to represent different rates of complaint generation from the crowd; note that it is possible for a single user to generate multiple complaints. The maximum values of the costless PCCA are given in Table 7.

Although a crowd is effectively limitless in many ways, a careful consideration shows that the crowd is not omniscient. Table 7 provides clear evidence that there are limits to the ability of a crowd. Two crucial factors drive these limitations. First, any individual member of the crowd has limited capacity for voting on assessments. Second, there is an inherent feedback loop in crowdvoting in the sense that a large crowd can generate a large number of complaints. The combination of these two factors effectively limits crowdvoting, an insight that any decision maker involved with crowdvoting should know.

7. Conclusions and Model Extensions

The focus of our paper is to find the most efficient crowdvoting mechanisms for a firm to achieve target service levels, viewed through an ease-of-implementation lens. We effectively consider four voting mechanisms, defined along two dimensions, $\{\text{static},\text{dynamic}\}\times\{\text{unit}-\text{weight},\text{accuracy}-\text{weight}\}$, as well as a static–dynamic hybrid mechanism for the majority-rule case. If voters are statistically independent, the dynamic accuracy-weighted mechanism, which utilizes sequential hypothesis testing and is analyzed in Section 5.2, is best (assuming $g \neq h$; c.f., Lemma 1). If voters are correlated, the method of Section 5.2 is not applicable, and the best approach is the static accuracy-weighted mechanism of Section 5.1. However, these accuracy-weighted voting mechanisms can be burdensome to implement because of their parameter-estimation requirements; a manager might choose to adopt a simpler majority-rule mechanism that is easier to implement,

Table 7. Maximum Values of Costless PCCA for $M=5{,}000{,}000$

	$N=50\%\times M$	$N=25\%\times M$	$N=10\%\times M$	$N=1\%\times M$
$(\mu_p,\mu_q)=(0.8,0.6)$	90.0%	96.7%	99.8%	≈ 100%

in which case the hybrid mechanism of Section 4.3 is best as it combines the benefits of dynamism and flexible consensus parameters. Furthermore, we found that positive correlation between voters is detrimental to system performance; this is due to diminished crowd diversity, which the literature has shown is a key factor for high-performance crowdsourcing. Our paper also discusses the matching of multiple voters with multiple assessments. We show that the crowd is not omniscient, and for high target values of the PCCA, expert (costly) firm employees are needed. We also characterize the limits of a costless crowdvoting system that does not use firm employees. We next discuss model extensions.

7.1. Dynamic Accuracies

Our paper assumes a static model, in which the distributions of voter accuracies, f_p and f_q, are stationary. However, in reality, a voter's ability to assess content might improve (or simply change) over time, which implies that the distributions f_p and f_q might also change over time. In this section, we discuss the impact dynamic accuracies would have on our models.

The majority-vote rules in Sections 4.1 and 4.2 are the least affected because only the means μ_p and μ_q of the distributions are needed. A time series model could potentially capture the relevant shifts in the crowd's abilities. For instance, if μ_p^t is the mean in period t, then the mean in period $t+1$ could be captured via exponential smoothing, $\mu_p^{t+1} = \alpha\mu_p^t + (1-\alpha)\tilde{\mu}$, where $\tilde{\mu}$ is a current observation (sample) of the crowd's mean ability to determine $T = 1$. More sophisticated time series models, such as Holt–Winter's method, can be used to capture trend and/or seasonality.

The accuracy-weighted voting rules of Sections 5.1 and 5.2 are more affected by dynamic crowd accuracies. The impact is reduced if the approximation in Proposition 5 is utilized because it only requires the means and standard deviations of the distributions f_p and f_q. The standard deviations can also be tracked using time series methods as discussed in the previous paragraph. However, if the exact probability $P(R(v) = T)$ is required, then the full distributions f_p and f_q must be updated. Because this is likely an arduous task with many technical difficulties, we do not recommend using the accuracy-weighted rule if the exact probability is desired and the accuracy distributions are not stationary. We similarly advise against the sequential hypothesis testing approach of Section 5.2 because the full distributions are utilized.

7.2. Nonbinary Voting Models

Our paper assumes a binary choice for each voter, corresponding to a binary underlying truth. In many applications, there might be more than two voting options. For instance, relabeling slightly, suppose a vote $v \in \{1, \ldots, L\}$ for $L \geq 3$. Arrow's (1950) impossibility theorem tells us that, when there are three or more voting options, no rank-ordering voting system can convert the individual voters' preferences into a consistent collective ranking (e.g., for a crowdvoting system). Therefore, generalizing our voting rules to three or more alternatives is ill advised in our opinion, and we suspect that most of our results are not extendable to this case.

Acknowledgments

The author thanks the editor-in-chief, associate editor, and three referees for very insightful suggestions that greatly improved the paper. Special thanks are due to the referee that recommended the hybrid strategy of Section 4.3.

Appendix A. Proofs

Proof of Proposition 1. We first consider the case in which $T = 1$, where $v_i = 1$ with probability p_i, and $v_i = -1$ with probability $1 - p_i$. The accuracy p_i has distribution f_p with mean $\mu_p \in [0, 1]$. The conditional moment-generating function of v_i is $E[e^{sv_i}|p_i] = p_i e^s + (1 - p_i)e^{-s}$. We next calculate the unconditional moment-generating function

$$E[e^{sv_i}] = \int_0^1 E\left[e^{sv_i}|p_i\right] f_p(p_i) dp_i = \int_0^1 \left(p_i e^s + (1 - p_i)e^{-s}\right) f_p(p_i) dp_i = \mu_p e^s + \left(1 - \mu_p\right)e^{-s},$$

which implies that $P(v_i = 1) = \mu_p$ and $P(v_i = -1) = 1 - \mu_p$ (i.e., a Rademacher distribution with parameter μ_p). If X is a binomial random variable with n trials and probability of success μ_p, then

$$\sum_{i=1}^{n} v_i = X - (n - X) = 2X - n.$$

This observation allows us to evaluate the probability that

$$\begin{aligned} P(R_u(v) = 1 | T = 1) &= P\left(\sum_{i=1}^{n} v_i \geq m_p\right) \\ &= P\left(X \geq \left\lceil \frac{n+m_p}{2} \right\rceil\right) \\ &= 1 - F_{\mu_p}\left(\left\lceil \frac{n+m_p}{2} \right\rceil - 1\right), \end{aligned}$$

where F_{μ_p} is the cumulative distribution function (CDF) of a binomial random variable with n trials and probability of success μ_p.

The analysis for the case in which $T = -1$ is similar. If Y is a binomial random variable with n trials and probability of success μ_q, then

$$\sum_{i=1}^{n} v_i = -Y + (n - Y) = n - 2Y,$$

and we can show that

$$\begin{aligned} P(R_u(v) = -1 | T = -1) &= P\left(\sum_{i=1}^{n} v_i \leq -m_q\right) \\ &= P\left(Y \geq \left\lceil \frac{n+m_q}{2} \right\rceil\right) \\ &= 1 - F_{\mu_q}\left(\left\lceil \frac{n+m_q}{2} \right\rceil - 1\right), \end{aligned}$$

where F_{μ_q} is the cumulative distribution function of a binomial random variable with n trials and probability of success μ_q. Using the distribution for T, we obtain an unconditional probability of correct crowd assessment:

$$P(R_u(v) = T) = \varrho\left(1 - F_{\mu_p}\left(\left\lceil \frac{n+m_p}{2} \right\rceil - 1\right)\right) + (1-\varrho)\left(1 - F_{\mu_q}\left(\left\lceil \frac{n+m_q}{2} \right\rceil - 1\right)\right). \quad \square$$

Proof of Proposition 2. Leveraging the proof of Proposition 1, the type I error can be written as

$$\begin{aligned} P(R_u(v) = -1 | T = 1) &= P\left(\sum_{i=1}^{n} v_i \leq -m_q\right) \\ &= P\left(X \leq \left\lfloor \frac{n-m_q}{2} \right\rfloor\right) \\ &= F_{\mu_p}\left(\left\lfloor \frac{n-m_q}{2} \right\rfloor\right). \end{aligned}$$

Likewise, the type II error can be written as

$$\begin{aligned} P(R_u(v) = 1 | T = -1) &= P\left(\sum_{i=1}^{n} v_i \geq m_p\right) \\ &= P\left(Y \leq \left\lfloor \frac{n-m_p}{2} \right\rfloor\right) \\ &= F_{\mu_q}\left(\left\lfloor \frac{n-m_p}{2} \right\rfloor\right). \quad \square \end{aligned}$$

Proof of Proposition 3. Conditional on $T = 1$, we have a biased random walk in which the step v_i obeys the following distribution: $P(v_i = 1) = \mu_p$ and $P(v_i = -1) = 1 - \mu_p$. Thus, standard probability theory (e.g., Steele 2000, page 6), determines the probability of hitting m_p before $-m_q$ as

$$\frac{\left(\frac{1-\mu_p}{\mu_p}\right)^{m_q} - 1}{\left(\frac{1-\mu_p}{\mu_p}\right)^{m_p+m_q} - 1}.$$

The required probability, conditional on $T = -1$, is symmetrical. □

Proof of Proposition 4. The proof is similar to that of Proposition 3 except it uses the expressions for expected hitting times of biased random walks (e.g., Steele 2000, p. 6) for each value of $T \in \{-1, 1\}$. □

Proof of Lemma 1. We suppress the subscript i. We first study the case in which $T = 1$. We consider the random variable $x = pv$, which has the distribution $P(x = p) = p$ and $P(x = -p) = 1 - p$, where p is drawn from distribution f_p. The conditional CDF of x is equal to

$$G(t|p) = \begin{cases} 0, & t < -p \\ 1 - p, & t \in [-p, p) \\ 1, & t \geq p. \end{cases}$$

We first consider the case in which $t \in [0, 1]$, where the CDF

$$G(t) = \int_0^1 G(t|p) f_p(p) dp = \int_0^t f_p(p) dp + \int_t^1 (1 - p) f_p(p) dp = 1 - \int_t^1 p f_p(p) dp,$$

which gives a density of $g(t) = t f_p(t)$ for $t \in [0, 1]$. Similarly, when $t \in [-1, 0)$, the CDF is

$$G(t) = \int_0^1 G(t|p) f_p(p) dp = \int_0^{-t} 0 f_p(p) dp + \int_{-t}^1 (1 - p) f_p(p) dp = \int_{-t}^1 (1 - p) f_p(p) dp,$$

which gives a density of $g(t) = (1 + t) f_p(-t)$ for $t \in [-1, 0)$.

If $T = -1$, then the random variable $x = qv$ has the distribution $P(x = -q) = q$ and $P(x = q) = 1 - q$, where q is drawn from distribution f_q. In this case, the conditional CDF of x is equal to

$$H(t|q) = \begin{cases} 0, & t < -q \\ q, & t \in [-q, q) \\ 1, & t \geq q. \end{cases}$$

Repeating the preceding analysis, we get $H(t) = \int_0^t f_q(q) dq + \int_t^1 q f_q(q) dq$ for $t \in [0, 1]$ and $H(t) = \int_{-t}^1 q f_q(q) dq$ for $t \in [-1, 0)$. The corresponding densities are $h(t) = (1 - t) f_q(t)$ for $t \in [0, 1]$ and $h(t) = -t f_q(-t)$ for $t \in [-1, 0)$. □

Proof of Proposition 5. Suppressing the index i, the mean of ρv, conditional on $T = 1$, can be calculated using the conditional density of Lemma 1:

$$E[\rho v | T = 1] = \int_{-1}^0 t(1 + t) f_p(-t) dt + \int_0^1 t^2 f_p(t) dt = (E[p^2] - E[p]) + E[p^2] = 2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p.$$

Similarly, the conditional second moment is

$$E\left[(\rho v)^2 \middle| T = 1\right] = \int_{-1}^0 t^2(1 + t) f_p(-t) dt + \int_0^1 t^3 f_p(t) dt = (E[p^2] - E[p^3]) + E[p^3] = \sigma_p^2 + \mu_p^2,$$

which gives the conditional standard deviation

$$\sqrt{E\left[(\rho v)^2 \middle| T = 1\right] - E[\rho v | T = 1]^2} = \sqrt{\sigma_p^2 + \mu_p^2 - \left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)^2}.$$

Repeating the analysis for the $T = -1$ case, we obtain

$$E[\rho v | T = -1] = \int_{-1}^0 -t^2 f_q(-t) dt + \int_0^1 t(f_q(t) - t f_q(t)) dt = -E[q^2] + (E[q] - E[q^2]) = \mu_q - 2\left(\sigma_q^2 + \mu_q^2\right)$$

and

$$E\left[(\rho v)^2 \middle| T = -1\right] = \int_{-1}^0 -t^3 f_q(-t) dt + \int_0^1 t^2(f_q(t) - t f_q(t)) dt = E[q^3] + (E[q^2] - E[q^3]) = \sigma_q^2 + \mu_q^2,$$

which gives the conditional standard deviation

$$\sqrt{E\left[(\rho v)^2 \middle| T = -1\right] - E[\rho v | T = -1]^2} = \sqrt{\sigma_q^2 + \mu_q^2 - \left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right)^2}.$$

Assumption 3 implies that the votes are independent. Conditional on $T = 1$, we use the CLT to approximate the random variable $\sum_{i=1}^{n} \rho_i v_i$ as a normal random variable with mean $n(2(\sigma_p^2 + \mu_p^2) - \mu_p)$ and standard deviation

$$\sqrt{n}\sqrt{\sigma_p^2 + \mu_p^2 - \left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)^2}.$$

Likewise, when $T = -1$, we approximate $\sum_{i=1}^{n} \rho_i v_i$ as a normal random variable with mean $n(\mu_q - 2(\sigma_q^2 + \mu_q^2))$ and standard deviation

$$\sqrt{n}\sqrt{\sigma_q^2 + \mu_q^2 - \left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right)^2}.$$

Finally,

$$\begin{aligned}
P\left(R_\rho(v) = T\right) &= \varrho P\left(R_\rho(v) = 1 | T = 1\right) + \left(1 - \varrho\right) P\left(R_\rho(v) = -1 | T = -1\right) \\
&= \varrho P\left(\sum_{i=1}^{n} \rho_i v_i \geq m_p | T = 1\right) + \left(1 - \varrho\right) P\left(\sum_{i=1}^{n} \rho_i v_i \leq -m_q | T = -1\right) \\
&\approx \varrho\left(1 - \Phi\left(\frac{m_p - n\left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)}{\sqrt{n}\sqrt{\sigma_p^2 + \mu_p^2 - \left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)^2}}\right)\right) + (1 - \varrho)\Phi\left(\frac{-m_q - n\left(\mu_q - 2\left(\sigma_q^2 + \mu_q^2\right)\right)}{\sqrt{n}\sqrt{\sigma_q^2 + \mu_q^2 - \left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right)^2}}\right) \\
&= \varrho\Phi\left(\frac{n\left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right) - m_p}{\sqrt{n}\sqrt{\sigma_p^2 + \mu_p^2 - \left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)^2}}\right) + (1 - \varrho)\Phi\left(\frac{n\left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right) - m_q}{\sqrt{n}\sqrt{\sigma_q^2 + \mu_q^2 - \left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right)^2}}\right). \quad \square
\end{aligned}$$

Proof of Proposition 6. Leveraging the proof of Proposition 5, the type I error can be written as

$$\begin{aligned}
P\left(R_\rho(v) = -1 | T = 1\right) &= P\left(\sum_{i=1}^{n} \rho_i v_i \leq -m_q | T = 1\right) \\
&\approx \Phi\left(\frac{-m_q - n\left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)}{\sqrt{n}\sqrt{\sigma_p^2 + \mu_p^2 - \left(2\left(\sigma_p^2 + \mu_p^2\right) - \mu_p\right)^2}}\right).
\end{aligned}$$

Likewise, the type II error can be written as

$$\begin{aligned}
P\left(R_\rho(v) = 1 | T = -1\right) &= P\left(\sum_{i=1}^{n} \rho_i v_i \geq m_p | T = -1\right) \\
&= 1 - \Phi\left(\frac{m_p - n\left(\mu_q - 2\left(\sigma_q^2 + \mu_q^2\right)\right)}{\sqrt{n}\sqrt{\sigma_q^2 + \mu_q^2 - \left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right)^2}}\right) \\
&= \Phi\left(\frac{-m_p - n\left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right)}{\sqrt{n}\sqrt{\sigma_q^2 + \mu_q^2 - \left(2\left(\sigma_q^2 + \mu_q^2\right) - \mu_q\right)^2}}\right). \quad \square
\end{aligned}$$

Proof of Proposition 7. Because $P(R(v) = 0) = 0$, analytically, we have that

$$\begin{aligned}
P(R(v) = T) &= \varrho P(R(v) = 1 | T = 1) + \left(1 - \varrho\right) P(R(v) = -1 | T = -1) \\
&\geq \varrho(1 - \epsilon) + \left(1 - \varrho\right)(1 - \varepsilon).
\end{aligned}$$

However, because there is an approximation associated with the Wald boundaries, we conservatively introduce the approximation. □

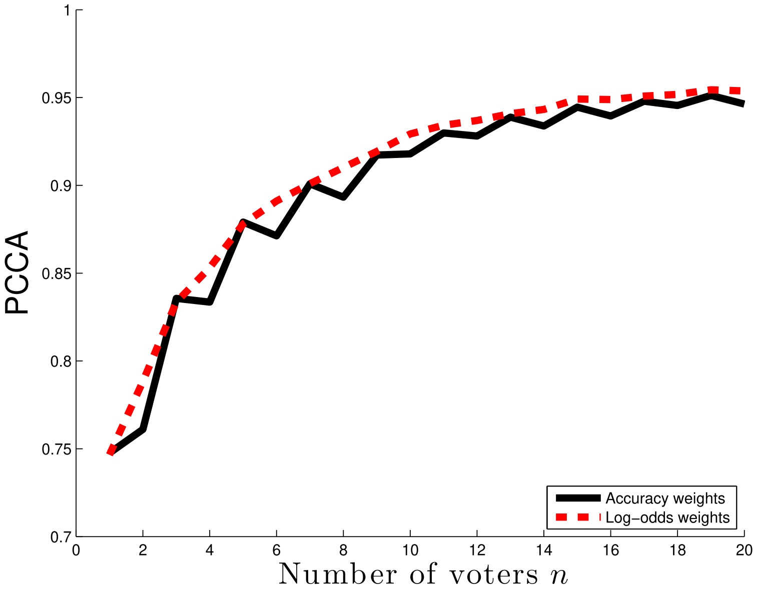
Proof of Proposition 8.

$$\begin{aligned} E[n] &= \varrho E[n|T=1] + (1-\varrho)E[n|T=-1] \\ &= \varrho \frac{E\big[\log(\Lambda(x_1,\dots,x_n))|T=1\big]}{E[z|T=1]} + (1-\varrho)\frac{E\big[\log(\Lambda(x_1,\dots,x_n))|T=-1\big]}{E[z|T=-1]} \quad \text{(Wald's Identity)} \\ &\approx \varrho\left(\frac{\epsilon\log(A)+(1-\epsilon)\log(B)}{E\Big[\log\Big(\frac{h(x)}{g(x)}\Big)|T=1\Big]}\right) + (1-\varrho)\left(\frac{(1-\varepsilon)\log(A)+\varepsilon\log(B)}{E\Big[\log\Big(\frac{h(x)}{g(x)}\Big)|T=-1\Big]}\right) \quad \text{(assuming overshoot is negligible)} \\ &= \varrho\left(\frac{\epsilon\log(A)+(1-\epsilon)\log(B)}{\int_{-1}^{1}\log\Big(\frac{h(t)}{g(t)}\Big)g(t)dt}\right) + (1-\varrho)\left(\frac{(1-\varepsilon)\log(A)+\varepsilon\log(B)}{\int_{-1}^{1}\log\Big(\frac{h(t)}{g(t)}\Big)h(t)dt}\right) \\ &= \varrho\left(\frac{\epsilon\log(\frac{1-\varepsilon}{\epsilon})+(1-\epsilon)\log(\frac{\varepsilon}{1-\epsilon})}{\int_{-1}^{1}\log\Big(\frac{h(t)}{g(t)}\Big)g(t)dt}\right) + (1-\varrho)\left(\frac{(1-\varepsilon)\log(\frac{1-\varepsilon}{\epsilon})+\varepsilon\log(\frac{\varepsilon}{1-\epsilon})}{\int_{-1}^{1}\log\Big(\frac{h(t)}{g(t)}\Big)h(t)dt}\right). \quad \square \end{aligned}$$

Appendix B. Accuracy Weights vs. Optimal Log-Odds Weights

Although the literature identifies the weights $w_i = \log(\frac{\rho_i}{1-\rho_i})$ as the optimal weights to maximize the PCCA for each n, we were unable to obtain the more insightful closed-form solutions for these weights. Fortunately, the suboptimality is minimal, as demonstrated in the following Monte Carlo simulation study. In particular, for each $n \in \{1,\dots,20\}$, we simulate 10,000 trials to evaluate the PCCA for both the accuracy weights $w_i = \rho_i$ and the log-odds weights $w_i = \log(\frac{\rho_i}{1-\rho_i})$, when f_p is a uniform distribution on $[0.7, 0.9]$ and f_q is a uniform distribution on $[0.5, 0.7]$. We plot both PCCAs in Figure B.1. Furthermore, the average ratio of the accuracy-weight PCCA to the log-odds-weight PCCA is 0.9915, and the ratio approaches one as n is increased. Finally, we observed similar results for different distributions f_p and f_q.

Figure B.1. The PCCAs for Accuracy and Log-Odds Weights



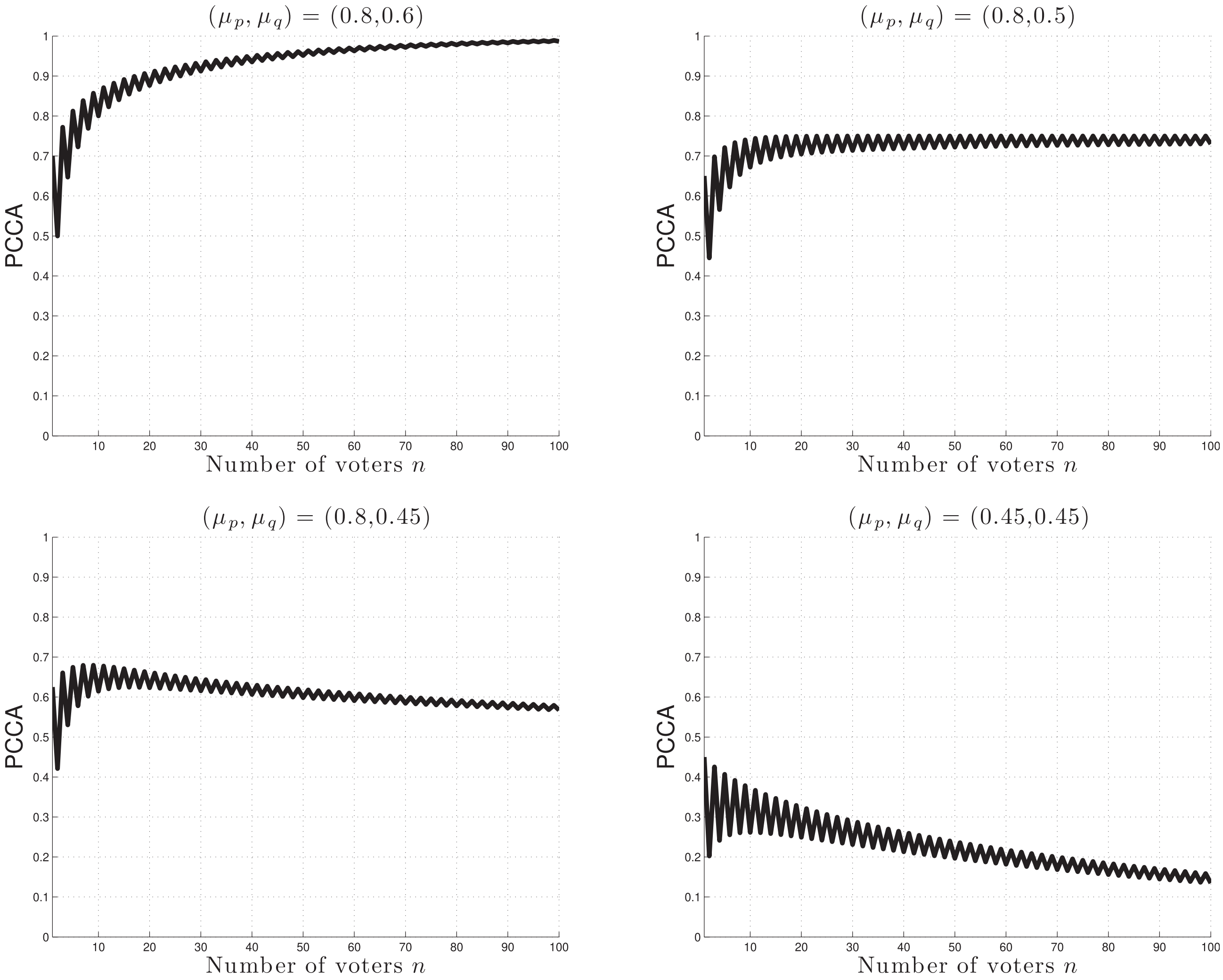
Appendix C. Nonmonotonicity of the PCCA

In this appendix, we demonstrate the cautionary result that the PCCA is not necessarily nondecreasing in the number of voters n. In Figure C.1, we plot the PCCA as a function of n for various sets of parameters. We fix $\varrho = 0.5$, but note that the following trends persist for other values of this parameter. We fix $m_p = m_q = 1$, but again note that the following patterns also persist for other values of these parameters, including values in which $m_p \neq m_q$. We see four possible trends in Figure C.1, one per panel of the figure, associated with the parameter values $(\mu_p, \mu_q) \in \{(0.8, 0.6), (0.8, 0.5), (0.8, 0.45), (0.45, 0.45)\}$. Each trend has a sawtooth pattern with local minima for every even n because of the potential of a tie rendering the crowd vote inconclusive; a similar observation, under a different voting model, appears in Dougherty and Edward (2009).

Therefore, we recommend avoiding an even number of voters under the unit-weight voting rule. We next explore each trend in more detail.

On the top left, we consider the case in which $\mu_p > 0.5$ and $\mu_q > 0.5$ and observe that the PCCA is effectively (ignoring even n) increasing in n to probability one, which agrees with the natural intuition that more voters is better. On the top right, we keep $\mu_p > 0.5$ but now set $\mu_q = 0.5$ and observe that the PCCA is again increasing in n, but to the value $\varrho + 0.5(1-\varrho)$; if $\mu_p = 0.5$ and $\mu_q > 0.5$, the limit would be $0.5\varrho + (1-\varrho)$. The next two plots show the cautionary result that the PCCA is not necessarily nondecreasing in the number of voters n. On the bottom left plot, we consider $\mu_p > 0.5$ and $\mu_q < 0.5$, and we observe that the PCCA is *unimodal* in n with a maximum of 68% when $n = 9$ voters. It is natural to ask when it is realistic that μ_q (or μ_p) is strictly less than 0.5. Possible drivers include bias and/or poor skill. For instance, voters might encounter assessments that they think are offensive, but if they have previously taken part in similar behavior, bias might drive them to vote that the assessments are not offensive. The motivation can be strategic or conscience driven. The point is that values of $\mu_q < 0.5$ (or $\mu_p < 0.5$) should not be ignored. Finally, on the bottom right plot, we consider $\mu_p < 0.5$ and $\mu_q < 0.5$, and observe that the PCCA is strictly decreasing in odd n. These asymptotic behaviors are easily explained by dissecting the result in Proposition 1. For instance, the binomial probability $F_{\mu_p}(\lceil \frac{n+m_p}{2} \rceil - 1)$ approaches its normal approximation $\Phi(\frac{n(\frac{1}{2}-\mu_p)+\frac{m_p}{2}-1}{\sqrt{n\mu_p(1-\mu_p)}})$ as $n \to \infty$. Consequently, if $\mu_p > 0.5$, this binomial probability goes to zero; if $\mu_p < 0.5$, the probability goes to one; and if $\mu_p = 0.5$, the probability goes to 0.5.

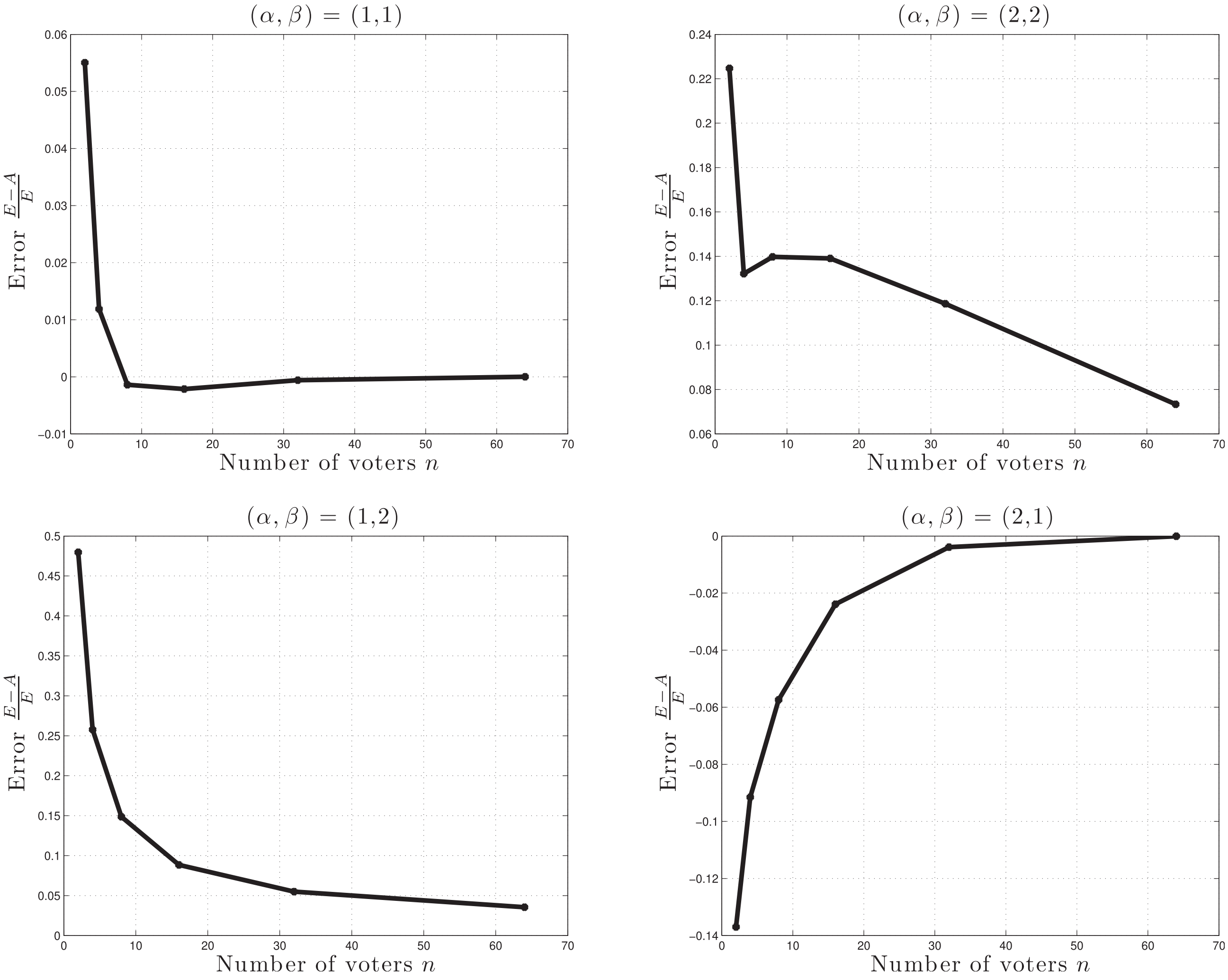
Figure C.1. The PCCA as a Function of n for $(m_p, m_q) = (1, 1)$



Appendix D. Evaluation of Approximation in Proposition 5

In this appendix, we test the quality of the approximation in Proposition 5. We focus on the $T = 1$ case because the $T = -1$ case is symmetric. In particular, we compare the approximate probability $A = \Phi(\frac{n(2(\sigma_p^2+\mu_p^2)-\mu_p)-m_p}{\sqrt{n}\sqrt{\sigma_p^2+\mu_p^2-(2(\sigma_p^2+\mu_p^2)-\mu_p)^2}})$ with the exact probability $E = P(R_\rho(v) = 1|T = 1)$, which is calculated using numerical convolution and integration with a discretization step size of $\delta = 0.0001$. We report the percentage error $\frac{E-A}{E}$ as a function of n for various assumptions on the underlying accuracy distribution f_p. We let f_p be a beta distribution with parameters α and β with mean $\mu_p = \alpha/(\alpha+\beta)$ and standard deviation $\sigma_p = \sqrt{\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))}$. We consider four sets of values for (α, β): (a) $(\alpha, \beta) = (1, 1)$ corresponding to a uniform distribution on $[0, 1]$, (b) $(\alpha, \beta) = (2, 2)$ corresponding to a symmetric unimodal distribution with a peak at 0.5, (c) $(\alpha, \beta) = (1, 2)$ corresponding to a distribution that is right skewed, and (d) $(\alpha, \beta) = (2, 1)$ corresponding to a distribution that is left skewed. Our results are plotted in Figure D.1 and can guide a manager to decide whether the approximation suffices for a given distribution f_p or whether a more sophisticated numerical implementation is needed. The approximation is best, not surprisingly, for the case in which f_p is a uniform distribution (top left panel). Because the normal approximation is symmetric, the right-skewed beta distribution leads to positive errors (bottom left panel) and the left-skewed beta distribution leads to negative errors (bottom right panel).

Figure D.1. The Approximation Errors $\frac{E-A}{E}$ as a Function of n



Endnote

[1] We focus on positive correlations. For large negative correlations and $n \geq 5$, the covariance matrix was not well defined (i.e., not positive definite); for small negative correlations, the PCCA was indistinguishable from the uncorrelated case.

References

Acemoglu D, Mostagir M, Ozdaglar A (2019) The economics of crowd organizations. Working paper, MIT, Cambridge, MA.

Amir O, Shahar Y, Gal Y, Ilany L (2013) On the verification complexity of group decision-making tasks. *Proc. First AAAI Conf. Human Comput. Crowdsourcing*.

Anderson A, Huttenlocher D, Kleinberg J, Leskovec J (2013) Steering user behavior with badges. *Proc. 22nd Internat. World Wide Web Conf.* (ACM, New York), 95–106.

Arrow K (1950) A difficulty in the concept of social welfare. *J. Political Econom.* 58(4):328–346.

Ben-Yashar R, Nitzan S (1997) The optimal decision rule for fixed-size committees in dichotomous choice situations: The general result. *Internat. Econom. Rev.* 38(1):175–186.

Berend D, Paroush J (1998) When is Condorcet's jury theorem valid? *Soc. Choice Welfare* 15(4):481–488.

Blake A (2018) A new study suggests fake news might have won Donald Trump the 2016 election. *Washington Post* (April 3), https://www.washingtonpost.com/news/the-fix/wp/2018/04/03/a-new-study-suggests-fake-news-might-have-won-donald-trump-the-2016-election/.

Boland P (1989) Majority systems and the Condorcet jury theorem. *Statistician* 38(3):181–189.

Budescu D, Chen E (2015) Identifying expertise to extract the wisdom of crowds. *Management Sci.* 61(2):267–280.

Burns C (2013) Steam users eclipse xBox Live, PSN still far and away tops. Accessed May 24, 2018, https://www.slashgear.com/steam-users-eclipse-xbox-live-psn-still-far-and-away-tops-30303588/.

Caldentey R, Araman V (2013) Crowdvoting the timing of new product introduction. Presentation at the 2013 Manufacturing & Service Oper. Management Conf. (INSEAD).

Chen A (2014) The laborers who keep dick pics and beheadings out of your Facebook feed. Accessed May 24, 2018, https://www.wired.com/2014/10/content-moderation/.

Dougherty K, Edward J (2009) Odd or even: Assembly size and majority rule. *J. Politics* 71(2):733–747.

Downs A (1957) *An Economic Theory of Democracy* (Harper & Row, New York).

Dwoskin E (2018) Facebook is rating the trustworthiness of its users on a scale from zero to 1. *Washington Post* (August 21), https://www.washingtonpost.com/technology/2018/08/21/facebook-is-rating-trustworthiness-its-users-scale-zero-one/.

Feddersen T, Pesendorfer W (1997) Voting behavior and information aggregation in elections with private information. *Econometrica* 65(5):1029–1058.

Feddersen T, Pesendorfer W (1998) Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *Amer. Political Sci. Rev.* 92(1):23–35.

Felder A (2009) Case study: The fivethirtyeight.com predictive model of the 2008 presidential election. *STATS: The Magazine for Students of Statistics* 50.

Gerling K, Gruner H, Kiel A, Schulte E (2005) Information acquisition and decision making in committees: a survey. *Eur. J. Political Econom.* 21(3):563–597.

Ghosh B, Sen P (1991) *Handbook of Sequential Analysis* (CRC Press, New York).

Goel V (2017) Facebook will add 3,000 monitors to screen offensive content. *Boston Globe* (May 3), https://www.bostonglobe.com/business/2017/05/03/facebook-will-add-monitors-screen-offensive-content/2xIu38AsEAMCCwzp1Wpp3K/story.html.

Grofman B, Owen G, Feld S (1983) Thirteen theorems in search of the truth. *Theory Decision* 15(3):261–278.

Karger D, Oh S, Shah D (2014) Budget-optimal task allocation for reliable crowdsourcing systems. *Oper. Res.* 62(1):1–24.

Ladha K (1992) The Condorcet jury theorem, free speech, and correlated votes. *Amer. J. Political Sci.* 36(3):617–634.

Ladha K, Miller G, Oppenheimer J (1996) Information aggregation by majority rule: Theory and experiments. Accessed May 24, 2018, http://www.gvpt.umd.edu/oppenheimer/research/jury.pdf.

Li H, Suen W (2009) Viewpoint: Decision-making in committees. *Canadian J. Econom.* 42(2):359–392.

Liu T, Yang J, Adamic L, Chen Y (2014) Crowdsourcing with all-pay auctions: A field experiment on taskcn. *Management Sci.* 60(8):2020–2030.

MacManus C (2012) League of Legends the world's "most played video game." Accessed May 24, 2018, https://www.cnet.com/news/league-of-legends-the-worlds-most-played-video-game/.

Marinesi S, Girotra K (2013) Information acquisition through customer voting systems. INSEAD Working Paper No. 2019/99/TOM, INSEAD, Fontainebleau, France.

Massoulié L, Xu K (2018) On the capacity of information processing systems. *Oper. Res.* 66(2):568–586.

Nelsen R (2007) *An Introduction to Copulas*, 2nd ed. (Springer, New York).

O'Brien S (2016) Facebook gets 1 million user violation reports a day. Accessed May 24, 2018, http://money.cnn.com/2016/03/12/technology/sxsw-2016-facebook-online-harassment/index.html.

Papanastasiou Y (2020) Fake news propagation and detection: A sequential model. *Management Sci.* 66(5):1826–1846.

Papanastasiou Y, Bimpikis K, Savva N (2018) Crowdsourcing exploration. *Management Sci.* 64(4):1727–1746.

Riker W, Ordeshook P (1968) A theory of the calculus of voting. *Amer. Political Sci. Rev.* 62(1):25–42.

Sah R, Stiglitz J (1986) The architecture of economic systems: Hierarchies and polyarchies. *Amer. Econom. Rev.* 76(4):716–727.

Sah R, Stiglitz J (1988) Committees, hierarchies and polyarchies. *Econom. J. (London)* 98:451–470.

Salant J, Curtis L (2012) Nate Silver-led statistics men crush pundits in election. *Bloomberg Businessweek* (November 7), http://www.businessweek.com/news/2012-11-07/nate-silver-led-statistics-men-crush-pundits-in-election.

Steele JM (2000) *Stochastic Calculus and Financial Applications* (Springer, New York).

Terwiesch C, Xu Y (2008) Innovation contests, open innovation, and multiagent problem solving. *Management Sci.* 54(9):1529–1543.

Thompson N (2018) Exclusive: Facebook opens up about false news. *Wired.com* (May 23), https://www.wired.com/story/exclusive-facebook-opens-up-about-false-news/.

Tiku N, Newton C (2015) Twitter CEO: "We suck at dealing with abuse." Accessed May 25, 2018, https://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the.

Wald A (1945) Sequential tests of statistical hypotheses. *Ann. Math. Statist.* 16(2):117–186.

Wald A, Wolfowitz J (1948) Optimum character of the sequential probability ratio test. *Ann. Math. Statist.* 19(3):326–339.