DRAFT: Do not share or distribute

Community Notes are Centrist, Not Consensus

Andreas Haupt

November 30, 2025

Abstract

We study the relationship between classic models of voting by committees and the bridging-based ranking algorithm used by X's Community Notes system for community fact-checking. We first represent monotonic, neutral and anonymous social choice rules via cutoffs for individual issues depending on association. We require a bridging axiom, and show that this demands a monotonicity condition from the dependence. With all but one issue fixed, the algorithm appears to estimate centrist preferences. For all but one voter fixed, it relates to a quadratic voting mechanism. As a purely statistical mechanism, it may lead to welfare loss as it does not incentivize users who are strongly correlated with many others to vote.

Keywords: voting by committees; Community Notes; Birdwatch; bridging algorithms; crowd-sourced fact-checking; generalized median voter schemes; latent-space models.

JEL codes: D71; D72; D83; C72.

1 Introduction

Online platforms increasingly rely on crowd-based fact-checking systems in which large numbers of users propose and rate annotations intended to correct or contextualize potentially misleading content. A prominent recent example is Community Notes on X (formerly Twitter), which allows registered contributors to attach short notes to posts and to rate others' notes for "helpfulness". Community Notes departs from earlier reputation- or majority-based systems by explicitly rewarding "bridging": its ranking algorithm is designed to surface notes that receive high helpfulness ratings from contributors who typically disagree with each other, rather than from a single partisan bloc.[1]

Community Notes can be described as follows. Contributors propose notes on posts. Other contributors then rate each note as helpful, somewhat helpful, or not helpful. These ratings form a sparse matrix of ordinal judgments, with rows corresponding to users and columns to notes. The platform fits a one-dimensional matrix factorization model in which each rater i and each note n is

[1] See the public documentation of Community Notes, which repeatedly emphasizes "diversity of perspectives" and describes a "bridging-based" ranking algorithm (Community Notes 2023), and the empirical and methodological discussion in Wojcik, Hilgard, Judd, Mocanu, Ragain, Fallin Hunzaker, et al. (2022).

associated with a scalar latent position θ_i and θ_n, together with rater and note intercepts α_i and β_n and a global intercept. The model treats each observed rating as a noisy realization of a latent utility or helpfulness index, often through a logistic or ordered-logit link:

$$\Pr\{y_{in} = 1\} \approx \sigma\big(\beta_0 + \alpha_i + \beta_n + \theta_i\theta_n\big),$$

where y_{in} is an indicator that rater i found note n helpful and σ is the logistic cdf.[2] The latent positions θ_i are interpreted as capturing a one-dimensional ideological or opinion spectrum along which raters differ, while the intercept β_n measures the extent to which note n is broadly helpful across that spectrum. Community Notes then ranks notes using a function of the estimated intercepts and their uncertainty, and surfaces a note when it attains a sufficiently high estimated intercept while also satisfying additional quality filters.

Bridging, in this context, refers to the requirement that a note be supported by raters who historically disagree with each other. The idealized slogan is that "people who have disagreed in the past need to agree now": a correction should be shown only when it is endorsed by contributors who usually take opposite sides on political or ideological disputes. The matrix factorization model attempts to operationalize this idea by assigning similar latent positions to raters who often agree on past notes and treating a note as more "bridging" when positive ratings come from raters whose latent positions are far apart. However, because ranking is based on the note intercept β_n and not on any explicit condition that *all* ideological camps endorse the note, the implemented rule is weaker than the slogan suggests. Community Notes is thus a concrete, large-scale mechanism that takes as input a high-dimensional pattern of approval and disagreement and outputs a subset of notes, but one whose normative properties are not yet understood in the language of social choice.

Classical social choice theory, by contrast, has emphasized aggregation rules that are independent of exogenous social organization. In the canonical model, there is a finite set of voters and a finite set of social alternatives; aggregation rules map profiles of individual preferences or approvals into collective choices, subject to axioms such as anonymity, neutrality, and various incentive constraints. A central insight of this literature is that, under mild conditions, such rules can be represented as "voting by committees" (Barberà, Sonnenschein, et al. 1991). In this framework, each alternative is chosen if and only if a certain family of coalitions—a committee—approves it. Committees treat all voters symmetrically, and the axioms do not reference parties, communities, or other group structure beyond the identities of individuals. The rules are designed to be robust to how voters happen to be clustered in the electorate.

Bridging mechanisms such as Community Notes break this symmetry by explicitly exploiting a partition of voters or a latent dimension summarizing disagreement. The social choice rule is no longer defined solely in terms of coalitions of individuals, but instead depends on where voters lie in an inferred ideological space. Conceptually, this raises new normative questions: when is it desirable

[2] See Wojcik, Hilgard, Judd, Mocanu, Ragain, Fallin Hunzaker, et al. (2022) and the platform's own description of the matrix-factorization scoring rule (Community Notes 2023). Closely related informal expositions are given by Warden (2024b) and Buterin (2023).

for a mechanism to privilege cross-group agreement, and what axioms characterize rules that do so in a principled way? How should one compare a bridging-based rule, which conditions on the geometry of disagreement, to a committee-based rule that is neutral to social organization?

The reliance on latent ideological positions connects Community Notes to a much older tradition in political science. Beginning with Poole and Rosenthal (1985a), the NOMINATE family of procedures uses spatial models and non-linear factor-analytic techniques to infer legislators' ideological locations from roll-call votes.[3] Subsequent work, including the IDEAL model of J. D. Clinton et al. (2004), developed Bayesian ideal-point estimators that view roll-call data as noisy signals of underlying ideological coordinates. These tools have been central to empirical work on polarization, party discipline, and representation, and are now standard in the measurement toolkit of political scientists.

Yet despite the sophistication of NOMINATE, IDEAL, and related ideal-point models, they have largely been used as *descriptive* tools. Their outputs serve to locate actors in an ideological space, to track polarization over time, or to estimate counterfactual voting outcomes under different coalition structures. They have not, with rare exceptions, been deployed as *normative design tools* for social choice mechanisms: the models summarize disagreement but do not prescribe how collective decisions should depend on it. Community Notes is exceptional in this respect. It imports a matrix-factorization / ideal-point apparatus into a mechanism that actually decides which crowd-sourced fact-checking notes are surfaced to users. Recent empirical work documents that these notes can significantly reduce engagement with and diffusion of false content on X (I. Slaughter et al. 2025a), but the underlying social choice rule remains conceptually opaque.

This paper develops a formal framework that connects bridging-based algorithms such as Community Notes to the theory of voting by committees. We formulate a model of approval aggregation in the style of Barberà, Sonnenschein, et al. (1991), extended to incorporate both anonymity and neutrality as well as a new "bridging" axiom. Voters submit approval ballots over a finite set of notes, and the mechanism selects a subset. In addition to the usual requirement that voters be treated symmetrically and notes be treated symmetrically, the bridging axiom requires that when a note enjoys broad support across a coarse partition of voters into "communities" or across a latent ideological dimension, it must be selected even if it fails to secure overwhelming support within any single community. We show in Theorem 4.1 that, on a rich domain of separable preferences, committee rules satisfying anonymity, neutrality, and the bridging axiom take a simple and intuitive form: they are equivalent to applying a voting-by-committees rule within each community, together with a cross-community quorum that requires support from voters who are far apart in the latent space.

We then characterize conditions under which the ranking induced by note intercepts can be rationalized by a bridging committee rule and conditions under which it cannot. This provides a precise sense in which Community Notes behaves "as if" it were implementing a particular bridging committee rule on the space of approval profiles induced by helpfulness ratings, while

[3] See Poole and Rosenthal (1985a) for the original spatial model and Poole and Rosenthal (1997a) for an overview; accessible introductions are provided by Voteview and related documentation (Lewis et al. 2025).

also highlighting the respects in which the platform's statistical objectives differ from standard normative criteria.

Third, we study axiomatic properties of anonymous, neutral, Paretian, and bridging social choice rules. We show that a class of mechanisms containing Community Notes satisfies this axiom.

We next analyze the marginal mechanism for a single voter and a single issue. For issues, the Community Notes rule infers the preferences of a centrist user. For users, the rule is connected to quadratic voting (Lalley and Weyl 2018): the effective weight of a rater's marginal report is proportional to the derivative of a quadratic loss, so that the mechanism aggregates information about intensity in a way reminiscent of voice-credit budgets, even though participants in Community Notes cast discrete approvals rather than buying votes.

Finally, we study welfare and selection-bias considerations when voting is costly and participation is endogenous. In particular, we identify conditions under which bridging reduces the influence of very strongly correlated groups, and conditions under which it instead overweights small but intensely engaged factions.

2 Literature Review

Voting by committees. Barberà, Sonnenschein, et al. (1991) study environments in which a finite set of voters must choose a subset of $K = \{1, \ldots, k\}$ objects, and preferences are separable across objects. They characterize the class of social choice functions that are strategy-proof, satisfy voter sovereignty, and respect natural symmetry requirements such as anonymity and neutrality. Any such rule can be represented as a family of committees, one for each object, together with associated quotas: an object is selected if and only if sufficiently many members of its committee vote for it. This representation exposes the underlying combinatorial structure of strategy-proof rules on binary outcome spaces and shows that non-dictatorial, strategy-proof mechanisms are available once preferences are restricted. Similarly, on single-peaked domains over a line, Moulin (1980) shows that any anonymous, strategy-proof and efficient rule is a generalized median voter scheme that selects a median among reported peaks and a finite set of "phantom" votes. Barberà, Gul, et al. (1993) and Barberà, Massó, et al. (1999) extend this perspective to multi-dimensional and lattice environments, characterizing maximal domains on which generalized median rules exhaust the set of strategy-proof and tops-only social choice functions.

Subsequent contributions apply and extend voting-by-committees and generalized median representations to richer economic settings. Barberà, Massó, et al. (2005) characterize strategy-proof and anonymous rules when the set of feasible subsets is constrained, showing that committee rules remain the relevant primitive once constraints are incorporated into the representation. Related work on strategy-proof correspondences formalizes how additional requirements such as Pareto efficiency, non-bossiness, or tops-only dependence further restrict the committee and generalized-median structures that can arise (Barberà, Dutta, et al. 2001; Barberà 2001). Taken together, this literature treats committees and generalized medians as canonical strategy-proof aggregators for binary or

coordinate-wise binary decisions.

Latent-dimension models of ideology. Our analysis also relates to the political science literature that infers low-dimensional ideological spaces from roll-call votes. Poole and Rosenthal (1985b) introduce NOMINATE, a spatial model in which each legislator is endowed with an ideal point in a Euclidean space and each roll-call vote corresponds to a pair of policy points. The probability that a legislator votes "yea" is a decreasing function of the distance between her ideal point and the "yea" point relative to the "nay" point. Estimation of this model yields coordinates for legislators and roll calls in one or two dimensions that explain the vast majority of observed voting behavior. Subsequent work, including the book-length treatment in Poole and Rosenthal (1997b) and the dynamic DW-NOMINATE procedure, refines this approach and documents the stability of a primary ideological dimension over long periods of US congressional history.

The IDEAL model of J. Clinton et al. (2004) reformulates roll-call scaling as a Bayesian item-response problem. Legislators' ideal points and roll calls' difficulty and discrimination parameters are treated as latent variables with prior distributions, and posterior inference is used to obtain uncertainty-quantified ideological placements. IDEAL and related models emphasize measurement properties and facilitate the incorporation of covariates or hierarchical structures, but—as with NOMINATE—take the voting rule as given and do not ask whether the roll-call procedure itself has desirable incentive or representation properties.

These latent-dimension models are thus descriptive tools: they map observed binary votes into a low-dimensional ideological space, which is then used to interpret patterns of agreement and disagreement. Our analysis borrows the idea that high-dimensional binary data can be usefully summarized by latent dimensions, but applies it normatively to the design of fact-checking mechanisms.

Community Notes and Deliberation. Twitter's Birdwatch program (now X's Community Notes) is the first large-scale deployment of community-based fact checking on a major social media platform. The official launch blog and subsequent product updates describe Birdwatch as a system in which enrolled users can flag potentially misleading posts, write short contextual notes, and rate others' notes for helpfulness (Twitter, Inc. 2021; Twitter, Inc. 2022). The platform later rebranded Birdwatch as Community Notes and open-sourced both its code and much of its data, including detailed documentation of the ranking algorithm that selects which notes are surfaced to the broader user base (X Corp. 2023b; X Corp. 2023a; Wikipedia contributors 2023). A central design feature is that a note is highlighted only if it receives high helpfulness ratings from users whose past ratings place them on different sides of a latent opinion spectrum, an explicitly bridging-based rather than majoritarian scoring rule.

A growing empirical literature analyzes the behavior of contributors and the properties of the Birdwatch / Community Notes system. Pröllochs (2022) provide an early holistic analysis of Birdwatch, documenting who writes notes, what kinds of content are fact-checked, and which notes are rated as helpful. They find that contributors tend to be highly engaged Twitter users and

that notes frequently target influential accounts. Saeed et al. (2022) compare crowd ratings on Birdwatch with expert fact-checks, showing that aggregated crowd judgments can approximate expert assessments, but also highlighting inconsistencies and challenges in obtaining reliable consensus. Allen et al. (2022) study partisanship in note evaluation, finding that contributors are more likely to flag and downrate content posted by political outgroups, which raises concerns about politically motivated behavior. Other work examines contributors' labor and participation patterns, noting that a relatively small subset of users produces most notes and ratings (Jones et al. 2022).

Several studies investigate the diffusion of content that has been community fact-checked. Drolsbach and Pröllochs (2022) analyze the virality of tweets that receive community fact-checks and show that crowd fact-checked misinformation is, on average, less viral than non-misleading tweets, in contrast to earlier findings for expert-fact-checked misinformation. Pilarski et al. (2023) compare the targeting behavior of Community Notes contributors with that of professional fact-checkers, finding systematic differences in which posts are selected for community versus expert scrutiny. Wirtschafter et al. (2023) document that, although only a minority of contributors ever write notes that are ultimately rated as helpful, the overall volume of helpful notes attached to tweets has grown substantially over time.

The most recent work evaluates the causal impact and perceived credibility of Community Notes. Drolsbach, Solovev, et al. (2024) present a large-scale survey experiment in which US respondents are exposed to misleading and non-misleading posts accompanied by either simple misinformation flags or actual Community Notes. They find that Community Notes are perceived as more trustworthy and more helpful for identifying misleading posts than generic flags, largely because of the contextual explanations they provide. Chuai, Tian, et al. (2024) exploit the roll-out of Community Notes as a natural experiment and report that the introduction of the feature did not substantially reduce aggregate engagement with misleading posts on X/Twitter. In contrast, Chuai, Pilarski, et al. (2024) use detailed repost time series and a difference-in-differences design to show that, conditional on a note being displayed, Community Notes substantially reduce subsequent reshares and increase the probability that misleading posts are deleted. These findings suggest that the system can be effective at the margin but may respond too slowly relative to the speed of content diffusion.

Taken together, this literature identifies key structural features of community-based fact checking: who participates in writing and rating notes, how notes are ranked and surfaced, and what behavioral outcomes are affected. However, existing studies treat the Community Notes algorithm as a given and do not model it as a social choice rule that maps binary helpfulness reports into display decisions. In particular, the algorithm's bridging requirement is described informally in terms of agreement between "diverse" raters, but it has not been embedded in an axiomatic framework or connected to the committee and generalized median characterizations discussed above. Our analysis fills this gap by interpreting Community Notes as a committee rule operating over latent communities of contributors and by asking which axioms force a fact-checking mechanism to take a bridging form.

Another broadly used mechanism for large-scale online deliberation is Polis (The Computational Democracy Project 2023). It aggregates opinions by clustering participants in a latent space and

visualizing the resulting structure. In a typical Polis conversation, participants respond to an open-ended prompt by submitting short statements and by voting "agree", "disagree", or "pass" on others' statements. The resulting vote matrix is analyzed using dimension reduction and clustering techniques such as principal components analysis and k-means, yielding a low-dimensional opinion space in which participants and statements are embedded (The Computational Democracy Project 2023). The platform surfaces statements that are widely supported within clusters, as well as those that attract support across clusters, and presents real-time visualizations of the opinion landscape to both participants and organizers. Case studies from civic and policy applications emphasize Polis' ability to identify points of common ground and to reveal distinct yet internally coherent opinion groups (UK Policy Lab 2022; Noema Magazine 2020).

From a social choice perspective, Polis can be interpreted as implementing voting by (learned) committees. Clusters of participants identified in the latent opinion space behave like committees whose internal aggregation rule is a simple majority over statements. A statement that is broadly supported within a cluster can be viewed as passing the committee for that group, while a statement that achieves support across multiple clusters corresponds to a bridging outcome that cuts across otherwise divergent communities. The platform's emphasis on "group-informed consensus"—statements that command support in more than one cluster—is conceptually close to the generalized median and committee rules discussed above, but the committees are discovered from data rather than imposed ex ante.

Outline. Section 3 introduces the formal model of approval aggregation with communities and latent ideological structure. **??** presents our axioms, including the bridging axioms, and states the main characterization and impossibility results. Section 5 considers the marginal mechanism for a single issue, connecting the rule to quadratic voting. Section 6 considers the marginal mechanism for a single voter, connecting the rule to centrism. Section 7 studies welfare and selection-bias considerations under costly participation.

3 Model

Environment and preferences. There is a finite set of voters $N = \{1, \dots, n\}$ and a finite set of binary issues $J = \{1, \dots, m\}$. For each issue $j \in J$, the society must decide whether to accept (1) or reject (0) the issue. An outcome is a vector $x = (x_j)_{j \in J} \in \{0,1\}^J$, where $x_j = 1$ means that issue j is accepted.

For $x \in \{0,1\}^J$ and $S \subseteq J$, let x_S denote the restriction of x to coordinates in S, and write $\mathbf{0}$ for the all-zero vector in $X = \{0,1\}^J$. For each $S \subseteq J$, write 1_S for the vector whose j-th component equals 1 if $j \in S$ and 0 otherwise.

Each voter $i \in N$ has a complete and transitive preference relation $\succsim_i$ on X; we write $\succ_i$ and $\sim_i$ for the strict and indifference parts, respectively. A *preference profile* is a tuple $\succsim = (\succsim_i)_{i \in N}$, and we let $\mathcal{D}_i$ denote the set of all admissible preferences for voter i and $\mathcal{D} = \prod_{i \in N} \mathcal{D}_i$ the set of

admissible profiles.

The paper focuses on a separable preference domain in the spirit of Barberà, Sonnenschein, et al. (1991).

Definition 3.1 (Separable preferences)**.** A preference is *separable* if for every j, and $\mathbf{v}_{-j}$

$$(1, \mathbf{v}_{-j}) \succsim_i (0, \mathbf{v}_{-j}) \Longleftrightarrow (1, \mathbf{0}_{-j}) \succsim_i (0, \mathbf{0}_{-j}).$$

Let $\mathcal{D}_i^{\text{sep}} \subseteq \mathcal{D}_i$ denote the set of separable preferences of voter i, and let $\mathcal{D}^{\text{sep}} = \prod_{i \in N} \mathcal{D}_i^{\text{sep}}$.

An equivalent formulation of separability is through utility functions. A preference relation $\succsim_i$ on X is separable across issues if there exist functions $u_{ij} : \{0,1\} \to \mathbb{R}$, $j \in J$, such that for all $x, y \in X$,

$$x \succsim_i y \quad \Longleftrightarrow \quad \sum_{j \in J} u_{ij}(x_j) \;\geq\; \sum_{j \in J} u_{ij}(y_j),$$

with strict inequality whenever $x \succ_i y$. It is sufficient for our purposes to represent preferences as $v = (v_{ij})_{i \in N, j \in J} \in \{0,1\}^{N \times J}$, where $v_{ij} = 1$ means that voter i (weakly) prefers acceptance to rejection of issue j. In the remainder of the paper we therefore work with vote profiles v as the primitive description of individual preferences, keeping in mind that they arise from an underlying separable preference domain $\mathcal{D}^{\text{sep}}$.

Social choice rules and axioms. A *social choice function* (SCF) on the separable domain is a mapping $f : \mathcal{D}^{\text{sep}} \longrightarrow X$, which assigns an outcome $f(\succsim)$ to every preference profile $\succsim \in \mathcal{D}^{\text{sep}}$. Using the approval representation, we can equivalently view an SCF as a mapping

$$F : \{0,1\}^{N \times J} \longrightarrow X,$$

where $F(v)$ is interpreted as the outcome chosen when voters cast issue-by-issue votes $v = (v_{ij})_{i \in N, j \in J}$. For $x \in X$, let x_j denote the j-th coordinate of x. For a vote profile v, write

$$S_j(v) \;=\; \{i \in N : v_{ij} = 1\}$$

for the set of supporters of issue j in v. We use the same letter f for the SCF on preferences and its vote-based representation F whenever no confusion can arise, and we write $f_j(v)$ for the j-th coordinate of $f(v)$. Anonymity and neutrality are defensible requirements for a mechanism, as they require that no voter or no issue shall be treated ex ante differently.

Definition 3.2 (Anonymity)**.** An SCF $F : \{0,1\}^{N \times J} \to X$ is *anonymous* if for every permutation σ of N and every vote profile v,

$$F(v) \;=\; F(\sigma \cdot v),$$

where $(\sigma \cdot v)_{ij} = v_{\sigma^{-1}(i), j}$ for all $i \in N$ and $j \in J$. A preference-based SCF $f : \mathcal{D}^{\text{sep}} \to X$ is anonymous if its induced vote-based representation F is anonymous.

Definition 3.3 (Neutrality)**.** Let $\Pi(J)$ denote the set of permutations of J. For $\pi \in \Pi(J)$ and $x \in X$, define $\pi x \in \{0,1\}^J$ by $(\pi x)_j = x_{\pi^{-1}(j)}$. For a vote profile v, define πv by $(\pi v)_{ij} = v_{i,\pi^{-1}(j)}$. An anonymous SCF $F : \{0,1\}^{N\times J} \to X$ is *neutral* if for every permutation $\pi \in \Pi(J)$ and every vote profile v,

$$F(\pi v) = \pi F(v).$$

A preference-based SCF f is neutral if its vote-based representation F is neutral.

We use a weak Paretian condition.

Definition 3.4 (Monotonicity)**.** A vote-based SCF $F : \{0,1\}^{N\times J} \to X$ is *issue-wise monotone* if for every issue $j \in J$ and every pair of vote profiles v, v' such that

$$v_{i\ell} = v'_{i\ell} \quad \text{for all } i \in N,\ \ell \neq j, \qquad v_{ij} \leq v'_{ij} \quad \text{for all } i \in N,$$

we have

$$f_j(v) = 1 \implies f_j(v') = 1.$$

That is, once an issue is accepted, increasing the set of its supporters (while holding all other votes fixed) cannot overturn acceptance.

Finally, we impose the usual strategy-proofness requirement on the separable preference domain.

Definition 3.5 (Strategy-proofness)**.** An SCF $f : \mathcal{D}^{\text{sep}} \to X$ is *strategy-proof* on $\mathcal{D}^{\text{sep}}$ if for every voter $i \in N$, every preference profile $\succsim = (\succsim_i, \succsim_{-i}) \in \mathcal{D}^{\text{sep}}$ and every alternative report $\widehat{\succsim}_i \in \mathcal{D}_i^{\text{sep}}$,

$$f(\succsim_i, \succsim_{-i}) \succsim_i f(\widehat{\succsim}_i, \succsim_{-i}) \quad \text{or} \quad f(\succsim_i, \succsim_{-i}) \sim_i f(\widehat{\succsim}_i, \succsim_{-i}).$$

That is, truthful reporting is a (weakly) dominant strategy for each voter when preferences are separable.

Voting by committees. A classic mechanism, and the main related mechanism for our study of Community Notes is *Voting By Committees* (Barberà, Sonnenschein, et al. 1991).

Definition 3.6 (Simple games)**.** A *simple game* on N is a map $W : 2^N \to \{0,1\}$ such that:

1. $W(\emptyset) = 0$ and $W(N) = 1$;
2. W is monotone: if $S \subseteq T \subseteq N$ and $W(S) = 1$, then $W(T) = 1$.

A coalition $S \subseteq N$ with $W(S) = 1$ is *winning*; a coalition with $W(S) = 0$ is *losing*. A winning coalition S is *minimal winning* if $W(T) = 0$ for every strict subset $T \subsetneq S$.

Definition 3.7 (Voting by committees)**.** A vote-based SCF $F : \{0,1\}^{N\times J} \to X$ is a *voting-by-committees rule* if there exists a family of simple games $\mathcal{W} = (W_j)_{j\in J}$ on N such that, for every vote profile v and every issue $j \in J$,

$$f_j(v) = 1 \quad \iff \quad W_j(S_j(v)) = 1.$$

Equivalently, each issue j is accepted if and only if the coalition of its supporters is winning according to W_j.

Under voting by committees, each issue j is associated with a collection $\mathcal{C}_j$ of minimal winning coalitions, which we refer to as the *committees for issue* j. Acceptance of issue j is then equivalent to the existence of a committee all of whose members vote in favor of the issue.

Formally, given a simple game W_j, let

$$\mathcal{C}_j \;=\; \{S \subseteq N : S \text{ is minimal winning for } W_j\}.$$

Then $S_j(v)$ is winning for W_j if and only if $S_j(v)$ contains some $C \in \mathcal{C}_j$ as a subset, by monotonicity. Thus Definition 3.7 can be equivalently rewritten as

$$f_j(v) = 1 \quad \Longleftrightarrow \quad \exists\, C \in \mathcal{C}_j \text{ such that } C \subseteq S_j(v).$$

Our axioms impose standard richness and incentive constraints on the SCF.

Definition 3.8 (Non-imposition)**.** An SCF $f : \mathcal{D}^{\text{sep}} \to X$ is *non-imposing* if for every outcome $x \in X$ there exists a preference profile $\succsim \in \mathcal{D}^{\text{sep}}$ such that $f(\succsim) = x$.

Non-imposition is the analogue of voter sovereignty in Barberà, Sonnenschein, et al. (1991). It ensures that the rule does not exclude any feasible outcome ex ante.

Proposition 3.9 (Barberà, Sonnenschein, et al. 1991)**.** *Let $f : \mathcal{D}^{sep} \to X$ be a social choice function whose vote-based representation F is non-imposing, and strategy-proof on $\mathcal{D}^{sep}$. Then there exists a family of simple games $\mathcal{W} = (W_j)_{j \in J}$ such that F is a voting-by-committees rule with respect to $\mathcal{W}$ in the sense of Definition 3.7. In particular, for each issue j there is a collection of committees $\mathcal{C}_j$ such that issue j is accepted if and only if some committee in $\mathcal{C}_j$ unanimously supports it.*

Example 3.10. As a first way to capture the idea of bridging is *consensus*. Call a *community structure* a partition $\mathcal{G} = \{G_1, \ldots, G_K\}$ of N into nonempty groups, with $K \geq 1$. The members of a given G_k are interpreted as sharing a common background, such as ideology, information sources, or expertise. Given a community structure $\mathcal{G}$, we view a coalition as *bridging* if it intersects all communities, $S \cap G_k \neq \emptyset$ for every $k \in \{1, \ldots, K\}$. Given a simple game W on N, a minimal winning coalition C for W is a *bridging committee* if C is a bridging coalition. A voting-by-committees rule F with committees $(\mathcal{C}_j)_{j \in J}$ is a *bridging voting-by-committees rule* (with respect to $\mathcal{G}$) if every committee $C \in \mathcal{C}_j$ is a bridging coalition for every issue $j \in J$.

We will, in the rest of this article, see that community notes does not have the properties of consensus.

Finally, we define our notion of bridging.

Definition 3.11 (Coalitions becoming more aligned)**.** Let $C \subseteq N$ and $S \subseteq J$. A profile R' is obtained from R by *making C more unanimous on S* if:

1. For every $k \in S$ and all $i, j \in C$,
$$x'_{ik} = x'_{jk},$$

2. For all (i, k) with either $i \notin C$ or $k \notin S$,
$$x'_{ik} = x_{ik}.$$

In this case we call R' a *bridging move* from R.

Definition 3.12 (Bridging order on profiles)**.** Define a binary relation $\preceq^{\text{br}}$ on profiles by $R \preceq^{\text{br}} R'$ if there exists a finite sequence
$$R = R^0, R^1, \ldots, R^L = R'$$
such that for each ℓ there is a coalition $C^\ell \subseteq \mathcal{N}$ and a set $S^\ell \subseteq \mathcal{K}$ with $R^{\ell+1}$ obtained from R^ℓ by making C^ℓ more unanimous on S^ℓ and satisfying
$$F_k(R^{\ell+1}) = F_k(R^\ell) \quad \text{for all } k \in S^\ell.$$

Definition 3.13 (Bridging axiom)**.** The rule F satisfies the *bridging axiom* if for every coalition $C \subseteq \mathcal{N}$, every nonempty $S \subseteq \mathcal{K}$, every profile R, and every profile R' obtained from R by making C more unanimous on S,
$$\Big(F_k(R') = F_k(R)\ \forall k \in S\Big) \implies \Big(F_\ell(R') = F_\ell(R)\ \forall \ell \in \mathcal{K} \setminus S\Big).$$

Definition 3.14 (Relatedness map and quota functions)**.** A *relatedness map* is a function
$$\rho : \mathcal{R} \to \mathbb{R}^d$$
for some $d \in \mathbb{N}$ that is anonymous and neutral, meaning that $\rho(R)$ depends on R only through anonymous and issue-symmetric statistics of agreement and disagreement between voters.

Given ρ, a family $(q_k)_{k \in \mathcal{K}}$ of functions
$$q_k : \mathbb{R}^d \to \{0, 1, \ldots, |\mathcal{N}|\}$$
are *quota functions* if for all $R \in \mathcal{R}$ and all $k \in \mathcal{K}$,
$$F_k(R) = 1 \quad \iff \quad \sum_{i \in \mathcal{N}} x_{ik} \ \geq\ q_k\big(\rho(R)\big).$$

Finally, we introduce Community Notes. Let Y be the note-by-rater matrix whose entry Y_{un} encodes rater u's evaluation of note n (helpful, somewhat helpful, not helpful). The production system fits a low-rank latent-factor model
$$\mathbb{E}[Y_{un} \mid \mu, \alpha_u, \beta_n, \theta_u, \theta_n] \ =\ g\big(\mu + \alpha_u + \beta_n + \theta_u \theta_n\big),$$

where g is a link function (logistic or ordered-logit), μ is a global intercept, α_u and β_n are user and note intercepts capturing overall harshness and helpfulness, and $\theta_u, \theta_n \in \mathbb{R}$ are one-dimensional latent factors that are interpreted as ideological positions.[4]

Given an estimated model $(\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\theta})$, the *Community Notes score* of a note n is essentially its fitted intercept $\hat{\beta}_n$. Notes with high $\hat{\beta}_n$ are those that, after controlling for the interaction $\theta_u \theta_n$, are judged helpful by a broad cross-section of raters; equivalently, they are notes that do well *even once polarization has been projected out.* The system declares a note n to be *Helpful* if $\hat{\beta}_n$ exceeds a global threshold and several additional guardrails are satisfied.[5]

4 Characterization

The next theorem shows that, under anonymity and neutrality, bridging is equivalent to a monotonicity condition on a suitably defined map from profiles into a Euclidean relatedness space.

Theorem 4.1 (Bridging characterization)**.** *Let F be an anonymous, neutral, monotonic, bridging social choice rule on $\mathcal{R}$. Then there exist:*

1. *a dimension $d \in \mathbb{N}$ and an anonymous and neutral* relatedness map $\rho : \mathcal{R} \to \mathbb{R}^d$*; and*
2. *for each issue $k \in \mathcal{K}$, a* quota function $q_k : \mathbb{R}^d \to \{0, 1, \ldots, |\mathcal{N}|\}$

such that for every profile $R \in \mathcal{R}$ and every issue $k \in \mathcal{K}$,

$$F_k(R) = 1 \quad \Longleftrightarrow \quad \sum_{i \in \mathcal{N}} x_{ik} \;\geq\; q_k\big(\rho(R)\big).$$

Moreover:

*(i) (*Bridging monotonicity*) For each k, the quota function q_k is monotone with respect to the bridging order $\preceq^{\mathrm{br}}$ in the sense that whenever $R' \succeq^{\mathrm{br}} R$ we have*

$$q_k\big(\rho(R')\big) \;\geq\; q_k\big(\rho(R)\big),$$

so that making coalitions more internally aligned on other issues can only (weakly) increase the quota they must meet on issue k.

(ii) Conversely, any rule that admits such a representation for some anonymous and neutral ρ and monotone quota functions $(q_k)_{k \in \mathcal{K}}$ satisfies the bridging axiom.

The vector $\rho(R) \in \mathbb{R}^d$ can be interpreted as an abstract representation of the *relatedness structure* of the profile: it summarizes, in an anonymous and neutral way, how often which sets of voters

[4] In practice, the implementation regularizes note intercepts and factors differently, but this distinction is not important here; see Community Notes (2024), Li et al. (2025), and Warden (2024c) for details.

[5] See Community Notes (2024) for the current thresholds and guardrails; similar descriptions appear in Chuai, Tian, et al. (2024) and I. Slaughter et al. (2025b).

agree or disagree across issues. The theorem states that (i) for each issue k the decision depends on R only via the total support on k and this relatedness vector, and (ii) bridging forces the quota on k to be monotone in ρ in the sense that a profile where voters are more internally aligned on other issues can never make acceptance of k easier.

Proof Sketch. The first step is to apply the committee representation of Barberà, Sonnenschein, et al. (1991) issue-by-issue: separability of preferences lets us treat each issue as a separate binary problem, while anonymity and neutrality across issues make the family of committees symmetric in voters and issues. On each issue k, any coalition $S \subseteq \mathcal{N}$ that can ever be decisive must be a member of some committee with an associated quota.

Next, use the group version of bridging. Fix an issue k and two profiles R, R' that agree on k and outside a coalition S, with $R' \succeq^{\mathrm{br}} R$. The bridging axiom implies that if S is not decisive for k at R it cannot be decisive at R'; making S more aligned on other issues cannot create new pivotal coalitions. Ordering coalitions by the profile of their within-coalition agreement on other issues and applying anonymity, one can show that for each issue the set of decisive coalitions must have a *threshold* structure: there is a minimal cardinality requirement that depends only on the anonymous relatedness profile of R. This defines the quota functions $(q_k)_{k\in\mathcal{K}}$ and the relatedness map ρ.

Finally, the equivalence between bridging and the monotonicity of $q_k \circ \rho$ is obtained by tracing how the set of pivotal coalitions changes as we move upwards in the order $\preceq^{\mathrm{br}}$.

Theorem 4.1 is deliberately agnostic about the dimension d and the specific choice of ρ. In applications, however, it is useful to have a more concrete representation. A natural candidate is the vector of agreement frequencies between all ordered pairs of voters, possibly compressed into a lower-dimensional embedding. □

Example 4.2. Community Notes is bridging according to Theorem 4.1. The relatedness map ρ is given by the empirical distribution of user factors $(\hat{\theta}_u)_{u\in\mathcal{N}}$ and their agreement patterns across issues; for a given rating matrix Y, the quota function q_k on issue k is implicit in the inequality $\hat{\beta}_k(Y) \geq t$, where t is the helpfulness threshold. Increasing cross-cutting agreement on other issues corresponds to shifting $\rho(Y)$ in the direction that makes it harder to explain the data by a single ideological factor, and the model then compensates by lowering intercepts for notes whose support is confined to more homogeneous groups.

Bridging as a constraint on pivotality. To connect this to the bridging axiom, fix some target note k and consider two rating matrices Y, Y' that agree on all ratings of k and on all ratings outside a coalition $S \subseteq \mathcal{N}$, but differ on a set of other notes $L \subseteq \mathcal{K}$ in such a way that raters in S become *more* internally aligned in Y' than in Y. Under mild regularity assumptions on the matrix factorization—uniqueness of the solution up to the usual sign indeterminacy, smooth dependence of $(\hat{\theta}, \hat{\beta})$ on the data, and the standard regularization that prefers to explain variation through θ rather than through β—one can show that:

- the spread of $(\hat{\theta}_u)_{u\in S}$ weakly decreases when passing from Y to Y'; and

- the fitted intercept $\hat{\beta}_k$ can only (weakly) decrease when the same pattern of ratings on k can be explained using a more homogeneous set of ideological factors.

Because the decision to display k is monotone in $\hat{\beta}_k$, this means that S cannot become newly pivotal on k by becoming more aligned on other notes: if k was not displayed at Y, then either it is still not displayed at Y', or it would also have been displayed at Y for some coalition that is at least as internally diverse. This is exactly the content of the bridging axiom.

Formally, this argument can be implemented by rewriting the factorization as a least-squares projection of the centered ratings onto the span of the user and note factors and showing that the residual component defining $\hat{\beta}_k$ is Schur-convex in the disagreement profile used to define the order $\preceq^{\mathrm{br}}$. The complete proof, which also accommodates missing ratings and the guardrail filters, is given in Appendix **??**.

Necessity, not sufficiency, for the bridging slogan. The characterization theorems above show that any Community Notes-style algorithm must satisfy the bridging axiom if it is to live up to the slogan that "people who often disagree need to agree" (Ovadya 2022a; Warden 2024c). Conversely, satisfying bridging is only a *necessary* condition: there are many relatedness-dependent quota rules that satisfy bridging but would not be considered good implementations of Community Notes. In particular, our theorems allow cutoffs that respond very weakly to cross-ideological agreement; Community Notes chooses a specific parametric form that makes the quota particularly sensitive to agreement between distant raters.

4.1 Dimensionality reduction and the choice of latent space

The Community Notes factorization uses a *single* latent ideological dimension: both users and notes are assigned positions $\theta_u, \theta_n \in \mathbb{R}$, and the interaction term is the product $\theta_u \theta_n$.[6] This mirrors the classic NOMINATE and IDEAL literatures in political science, where much of the variation in roll-call votes is captured by a single left–right axis.(Poole and Rosenthal 1985b; Poole and Rosenthal 1997a; J. D. Clinton et al. 2004)

From the standpoint of Theorem **??**, this amounts to choosing $D = 1$ and treating $\Delta(R)$ as a summary of agreement rates at various distances along that single axis. This has two important consequences.

First, it makes the bridging property *interpretable*: distance along the latent axis can be read, at least in the U.S. context, as an ideological or partisan distance. A note with a high intercept is then one that garners support from raters across that axis, which is precisely the kind of bridging we care about in political fact-checking.[7]

[6] Discussion of the one-dimensional implementation and proposals for multi-dimensional variants can be found in Warden (2024c) and Warden (2024a). Empirically, legislative roll call data are famously close to one-dimensional in many contexts (Poole and Rosenthal 1985b; Poole and Rosenthal 1997a; J. D. Clinton et al. 2004).

[7] See I. Slaughter et al. (2025b) for evidence that such notes reduce engagement with misinformation, and Chuai, Tian, et al. (2024) for complementary results on the impact of the system.

Second, compressing relatedness into a single dimension inevitably discards information. In environments where disagreement is multi-dimensional—for instance when debates simultaneously involve economic policy, cultural issues, and attitudes toward institutions—the leading latent factor may mix these cleavages in hard-to-interpret ways.[8] From the perspective of Theorem **??**, the system is still a bridging rule (it uses a relatedness-dependent quota that is monotone in agreement between distant raters), but the normative content of "distance" becomes murkier.

A richer representation with $D > 1$ would allow the algorithm to recognize multiple, potentially orthogonal, axes of disagreement. Our characterization theorem continues to apply: the relatedness map $\Delta(R)$ would track agreement rates at different distances in $\mathbb{R}^D$, and the quota functions (q_k) would be required to be monotone in the corresponding partial order. What changes is not the axiomatic structure, but the semantic interpretation of $\Delta(R)$ and the extent to which the bridging dimension(s) align with the salient social cleavages.

4.2 Polis as voting by learned committees

Polis provides a useful contrast to Community Notes. In its standard deployment, participants respond to a collection of statements ("prompts") by voting agree, disagree, or pass. The platform uses these responses to embed participants into a low-dimensional space and then clusters them into groups, which we denote by $\mathcal{G} = \{G_1, \dots, G_m\}$.[9] Each cluster G_j acts as a *learned committee*: for each statement ℓ, the system computes the share of participants in each G_j that vote "agree" and highlights those statements that reach high support in *all* clusters or in all but a small number of clusters.

In the language of Barberà, Sonnenschein, et al. (1991), this is very close to classical voting by committees. The clusters G_j are the committees, the internal decision rule of each G_j is simple majority, and the platform displays statements that belong to the intersection of the committees' winning sets. If the clustering is held fixed, the resulting rule is a standard issue-by-issue committee rule.

When we take seriously that the clusters are learned from the very same response matrix that is being aggregated, Polis can still be seen as an instance of Theorem 4.1. The relatedness map ρ assigns to each profile R the vector of cluster-level support rates for each statement, and the quotas q_k require a high level of support in each cluster for a statement k to be surfaced as consensus. The group version of the bridging axiom holds at the level of clusters: making a cluster G_j more internally homogeneous on other statements cannot increase its pivotal power on statement k, because the final decision depends only on the cluster-level majority on k.

From the standpoint of this paper, Polis is therefore *closer* to the classical voting-by-committees paradigm than Community Notes. The committees are explicit (the clusters), their internal rules

[8] This concern is emphasized in Warden (2024a) and in recent discussions of the international deployment of Community Notes, where the learned axis may capture, for example, an anti-elite dimension rather than a standard left–right one.

[9] See Small et al. (2021) for a detailed account of Polis as used in the vTaiwan process and other deliberative settings, and Salganik and Levy (2015) for the closely related "wiki survey" methodology on which Polis builds.

are simple majorities, and the system effectively applies an intersection-of-majorities rule across committees. Community Notes, in contrast, departs more radically from committee separability: there are no fixed committees, and the "committees" implicit in the factorization are changing, soft coalitions defined by latent positions and agreement patterns.

4.3 The strong bridging slogan and why Community Notes violates it

The bridging axiom is intentionally weak: it forbids certain kinds of cross-issue bossiness but does not insist that every accepted note must command support from voters who "often disagree." The public rhetoric around Community Notes, however, sometimes suggests a much stronger property: that a note is accepted only if such bridging occurs.

A strong bridging condition. To make this precise, fix a profile R and define a measure of historical disagreement between voters i and j,

$$d_{ij}(R) \;=\; \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} \mathbf{1}\{x_{ik} \neq x_{jk}\},$$

which counts how often i and j disagree across issues. Let $\bar{d}(R)$ be the cross-sectional mean of $d_{ij}(R)$ across unordered pairs $\{i,j\}$. Fix a disagreement threshold $\tau > 0$ and say that a pair (i,j) *often disagrees* at R if $d_{ij}(R) \geq \bar{d}(R) + \tau$.

We can now formalize the strong slogan as follows.

Definition 4.3 (Strong bridging)**.** A decision rule F satisfies *strong bridging* at threshold τ if for every profile R and issue k with $F_k(R) = 1$ there exists a pair (i,j) that often disagrees at R and such that $x_{ik} = x_{jk} = 1$. In words, whenever an issue or note is accepted, at least one pair of voters that frequently disagrees across issues must agree in supporting it.

Strong bridging prescribes an *existence* condition on the support coalition: every accepted note must contain a pair of historically distant voters that both endorse it. By contrast, the (weak) bridging axiom in Section 3 is a *monotonicity* condition: it restricts how pivotal coalitions can change when voters become more similar on other issues.

In terms of our characterization, strong bridging would require that the quota functions (q_k) depend only on statistics of support by high-disagreement pairs and that these statistics satisfy a hard lower bound whenever $F_k(R) = 1$. This is strictly stronger than the monotonicity requirement of Theorem 4.1.

Community Notes satisfies bridging but not strong bridging. The Community Notes algorithm does not satisfy Definition 4.3. The one-dimensional latent factor θ used in the factorization is estimated from the entire rating matrix Y and captures the dominant axis of disagreement in the data;[10] the intercept $\hat{\beta}_n$ for a note n measures how helpful the note is *after* projecting out this

[10] See Poole and Rosenthal (1985b), Poole and Rosenthal (1997a), and J. D. Clinton et al. (2004) for analogous results in roll-call data, and Warden (2024a) for a discussion of multi-dimensional extensions.

axis.(Community Notes 2024; Li et al. 2025; Warden 2024c) A note can therefore achieve a high intercept in at least two conceptually distinct ways:

1. it is rated helpful by raters spread across the entire ideological spectrum (including very distant pairs), or

2. it is rated helpful by a large collection of moderately diverse raters whose latent positions cover most of the mass of the θ distribution, even if the most extreme raters on each side do not endorse it.

In the second case, the set of raters who endorse the note may contain *no* pair (i, j) with $d_{ij}(R) \geq \bar{d}(R) + \tau$ for a reasonably high threshold τ. For example, suppose latent positions are approximately uniformly distributed on $[-2, 2]$ but almost all highly engaged raters lie in $[-1.5, -0.5] \cup [0.5, 1.5]$. Consider a note that is rated helpful by many raters in $[-1.5, -0.5]$ and $[0.5, 1.5]$ but almost no raters in the extreme tails $[-2, -1.5)$ and $(1.5, 2]$. It is easy to construct a profile in which:

- the Community Notes intercept $\hat{\beta}_n$ is high (because the note performs well across the main mass of the distribution of θ and residual variation is small), so n is displayed; but

- for every pair (i, j) that often disagrees in the sense of Definition 4.3 (say, one in the far left tail and one in the far right tail), at least one of i or j does *not* rate the note as helpful.

In such a profile F satisfies the bridging axiom—no coalition becomes newly pivotal on other notes by becoming more internally aligned in its off-note ratings—but it fails strong bridging because the accepted note does not contain a pair of voters that "often disagree" according to $d_{ij}(R)$.

This observation is consistent with empirical analyses of Community Notes. Studies find that notes deemed helpful tend to attract support from users with diverse estimated latent positions and that such notes reduce engagement with misinformation,(Chuai, Tian, et al. 2024; I. Slaughter et al. 2025b) but they do not claim that every helpful note is supported by the most polarized users on opposite ends of the spectrum. From the perspective of our characterization, Community Notes is a particular relatedness-dependent quota rule that satisfies the bridging axiom but falls short of the very demanding strong bridging condition.

5 A Single Voter

5.1 A Stylized Community Notes Estimator

We consider a fixed, finite set of notes $J = \{1, \ldots, m\}$ and a sequence of rating populations $N_n = \{1, \ldots, n\}$ with $n \to \infty$. Voter i^* is fixed once and for all and, without loss, we take $i^* = 1$. For each n, let

$$Y^{(n)} = \left(y_{ij}^{(n)}\right)_{i \in N_n,\, j \in J}$$

denote the (partially observed) rating matrix, where $y_{ij}^{(n)} \in \{-1, 0, 1\}$ is voter i's rating of note j: $+1$ for "helpful", -1 for "not helpful", and 0 for "no rating". Let $E_n \subseteq N_n \times J$ denote the set of observed pairs (i, j) with $y_{ij}^{(n)} \neq 0$.

Following the matrix-factorization description of Community Notes, the algorithm associates to each voter i a latent polarity $\theta_i \in \mathbb{R}$ and a voter intercept $\alpha_i \in \mathbb{R}$, and to each note j a note polarity loading $\beta_j \in \mathbb{R}$ and a note intercept $\mu_j \in \mathbb{R}$.Warden (2024d) models the fitted score for (i, j) as

$$\hat{y}_{ij} = \mu_j + \alpha_i + \beta_j \theta_i,$$

which is precisely the one-dimensional matrix-factorization model discussed in the Community Notes documentation and early empirical work on the system.(**CommunityNotesGuide**; B. Slaughter et al. 2025)

Let $\mu = (\mu_j)_{j \in J}$, $\alpha = (\alpha_i)_{i \in N_n}$, $\beta = (\beta_j)_{j \in J}$, and $\theta = (\theta_i)_{i \in N_n}$. Collect these in a parameter vector

$$\psi = (\mu, \alpha, \beta, \theta) \in \mathbb{R}^{2m+2n}.$$

Given $Y^{(n)}$ and E_n, the stylized Community Notes estimator is the penalized least-squares (equivalently, Gaussian-maximum-likelihood with Gaussian priors) problem

$$Q_n(\psi) \;:=\; \frac{1}{|E_n|} \sum_{(i,j) \in E_n} \left(y_{ij}^{(n)} - \mu_j - \alpha_i - \beta_j \theta_i\right)^2 \;+\; \lambda \left(\|\mu\|_2^2 + \|\alpha\|_2^2 + \|\beta\|_2^2 + \|\theta\|_2^2\right), \tag{1}$$

where $\lambda > 0$ is a regularization hyperparameter. The estimated parameters are

$$\hat{\psi}_n = \left(\hat{\mu}^{(n)}, \hat{\alpha}^{(n)}, \hat{\beta}^{(n)}, \hat{\theta}^{(n)}\right) \in \arg\min_{\psi} Q_n(\psi).$$

Empirically, the Community Notes system uses a "common-ground" or helpfulness dimension based on the note intercepts: a note with large positive intercept μ_j is predicted to be rated helpful by users across the latent polarity spectrum.Warden (2024d) In our stylized model, we therefore take the estimated helpfulness score of note j to be

$$h_j(\hat{\psi}_n) := \hat{\mu}_j^{(n)}$$

and say that note j is selected for display if $h_j(\hat{\psi}_n)$ exceeds a fixed threshold $\tau \in \mathbb{R}$. The resulting set of selected notes is

$$S_n := S(\hat{\psi}_n) := \left\{ j \in J : h_j(\hat{\psi}_n) \geq \tau \right\}.$$

Throughout this section we fix the ratings of all voters $i \neq i^*$ and the set of notes J, and consider counterfactual modifications of voter i^*'s row of the rating matrix while holding all other rows fixed.

5.2 Vanishing marginal influence of a single voter

We now formalize the sense in which the marginal influence of an individual voter on parameter estimates, and hence on the selected set of notes, vanishes as n becomes large.

For each n, consider two rating profiles:

- the baseline profile $Y^{(n)}$ with estimator $\hat{\psi}_n$ and selected set S_n;
- a deviating profile $\tilde{Y}^{(n)}$ that coincides with $Y^{(n)}$ for all voters $i \neq i^*$ but may differ arbitrarily in the row of i^*, with corresponding estimator $\tilde{\psi}_n$ and selected set $\tilde{S}_n$.

We write $\hat{\psi}_n = (\hat{\mu}^{(n)}, \hat{\alpha}^{(n)}, \hat{\beta}^{(n)}, \hat{\theta}^{(n)})$ and $\tilde{\psi}_n = (\tilde{\mu}^{(n)}, \tilde{\alpha}^{(n)}, \tilde{\beta}^{(n)}, \tilde{\theta}^{(n)})$. Since selection depends only on the note-side parameters, it is convenient to write

$$\varphi_n := (\hat{\mu}^{(n)}, \hat{\beta}^{(n)}) \in \mathbb{R}^{2m}, \qquad \tilde{\varphi}_n := (\tilde{\mu}^{(n)}, \tilde{\beta}^{(n)}).$$

We impose a mild regularity condition.

Assumption 5.1 (Regularity and bounded exposure). For each n:

1. Each voter rates at most $R_{\max}$ notes, i.e. $|\{j : (i,j) \in E_n\}| \leq R_{\max}$ for all i. In particular, the focal voter i^* contributes at most $R_{\max}$ entries to E_n.
2. Ratings are uniformly bounded: $|y_{ij}^{(n)}| \leq 1$ for all (i,j).
3. The per-edge loss
$$\ell_{ij}(\psi) := \left(y_{ij}^{(n)} - \mu_j - \alpha_i - \beta_j \theta_i\right)^2$$
has gradient with respect to $\varphi = (\mu, \beta)$ uniformly bounded on a ball that contains all minimizers of Q_n and of the perturbed objective defined by $\tilde{Y}^{(n)}$. That is, there exists $L < \infty$ such that
$$\|\nabla_\varphi \ell_{ij}(\psi)\|_2 \leq L$$
for all (i,j) and all ψ in that ball.
4. The objective Q_n is κ-strongly convex in φ uniformly in n on that ball: for all ψ, ψ' with note-side components φ, φ',
$$Q_n(\psi') \;\geq\; Q_n(\psi) + \nabla_\varphi Q_n(\psi)^\top (\varphi' - \varphi) + \frac{\kappa}{2}\|\varphi' - \varphi\|_2^2.$$

Assumption 5.1 is standard in the analysis of ridge-regularized matrix factorization and is satisfied, for example, if we restrict attention to a compact parameter set and use an ℓ_2 penalty with $\lambda > 0$ in (1) (Amatriain et al. 2011).

Proposition 5.2 (Asymptotic negligibility of a single voter)**.** *Suppose Assumption 5.1 holds and that the average number of ratings per voter converges to a positive constant,*

$$\frac{|E_n|}{n} \;\to\; \bar{R} \in (0,\infty) \quad \text{as } n \to \infty.$$

Then there exists a constant $C > 0$, independent of n, such that for all n,

$$\big\|\tilde{\varphi}_n - \varphi_n\big\|_2 \;\le\; \frac{C}{n}.$$

In particular, the induced change in helpfulness scores is uniformly $O(1/n)$:

$$\max_{j\in J}\big|h_j(\tilde{\psi}_n) - h_j(\hat{\psi}_n)\big| = \max_{j\in J}\big|\tilde{\mu}_j^{(n)} - \hat{\mu}_j^{(n)}\big| \;\le\; \frac{C}{n}.$$

Proof sketch. Let Q_n and $\tilde{Q}_n$ denote the objectives defined by $Y^{(n)}$ and $\tilde{Y}^{(n)}$, respectively. By construction, the two objectives differ only on those terms involving the focal voter i^*, of which there are at most $R_{\max}$. Write $\nabla_\varphi Q_n$ for the gradient of Q_n with respect to $\varphi = (\mu, \beta)$. By Assumption 5.1(3),

$$\sup_{\psi}\big\|\nabla_\varphi Q_n(\psi) - \nabla_\varphi \tilde{Q}_n(\psi)\big\|_2 \;\le\; \frac{2LR_{\max}}{|E_n|} \;=\; O\!\left(\frac{1}{n}\right).$$

Strong convexity of Q_n in φ implies a standard Lipschitz-continuity property of the argmin mapping: if f and g are κ-strongly convex functions with minimizers x_f^* and x_g^*, then

$$\|x_f^* - x_g^*\|_2 \;\le\; \frac{1}{\kappa}\sup_x\big\|\nabla f(x) - \nabla g(x)\big\|_2.$$

Applying this inequality to Q_n and $\tilde{Q}_n$ with the bound above yields

$$\big\|\tilde{\varphi}_n - \varphi_n\big\|_2 \;\le\; \frac{2LR_{\max}}{\kappa\,|E_n|} \;=\; O\!\left(\frac{1}{n}\right),$$

with a constant C that depends only on $(L, R_{\max}, \kappa, \bar{R})$. The bound on helpfulness scores follows because $h_j(\psi) = \mu_j$ projects φ onto coordinates of unit Lipschitz constant. Full details, including the extension to multiple latent factors, are deferred to the appendix. □

To translate this into a statement about the selected set S_n, we impose a simple margin condition at the limit.

Assumption 5.3 (Selection margin). There exists a parameter vector ψ^∞ and $\delta > 0$ such that the note intercepts at ψ^∞ are separated from the threshold:

$$\min_{j\in J}\big|\mu_j^\infty - \tau\big| \;\ge\; 2\delta, \qquad \text{where} \quad \mu^\infty = (\mu_j^\infty)_{j\in J}.$$

Moreover, the estimators converge in probability to ψ^{∞} under the baseline ratings.

Assumption 5.3 says that in the large-n limit, no note lies exactly at the selection threshold. This is generic in smooth models.(B. Slaughter et al. 2025)

Corollary 5.4 (Eventual agreement of selected sets)**.** *Suppose Assumptions 5.1 and 5.3 hold. Then there exists N_0 such that for all $n \geq N_0$ and for any deviation of voter i^*'s ratings,*

$$\tilde{S}_n = S_n.$$

More generally, for finite n the symmetric difference between the selected sets is empty whenever $C/n < \delta$, and is uniformly bounded in size whenever the margin condition holds with heterogeneous gaps.

Proof sketch. By Assumption 5.3, there exists N_0 such that for all $n \geq N_0$ and all $j \in J$,

$$\left|\hat{\mu}_j^{(n)} - \mu_j^{\infty}\right| \leq \delta \quad \text{with high probability.}$$

Combined with Proposition 5.2, we obtain for $n \geq N_0$:

$$\left|\tilde{\mu}_j^{(n)} - \hat{\mu}_j^{(n)}\right| \leq \frac{C}{n} < \delta.$$

Thus both $\hat{\mu}_j^{(n)}$ and $\tilde{\mu}_j^{(n)}$ lie on the same side of the threshold τ for every note j, so $j \in S_n$ if and only if $j \in \tilde{S}_n$. The more general statement follows from the same argument applied note-by-note with note-specific margins. Again, details are deferred to the appendix. □

Corollary 5.4 gives a precise sense in which, in a large electorate, an individual voter's deviation has vanishing impact on the selected set of Community Notes: their ratings perturb the note-level parameters by $O(1/n)$ and, under a generic margin condition, eventually leave the selected set unchanged. This echoes classical large-population results for quadratic voting and related mechanisms, where each agent behaves approximately as a price-taker.(Lalley and Weyl 2018; Georgescu et al. 2024)

5.3 A quadratic-voting analogy with exogenous polarity

We now connect the focal voter's behavior to quadratic voting when their latent polarity is held fixed.

Quadratic voting (QV) allocates to each voter an artificial budget of "voice credits" that they can spend across issues.Lalley and Weyl (2018) and Casella and Sanchez (2018) For a set of binary issues $K = \{1, \ldots, m\}$, a voter chooses a vector of votes $v = (v_k)_{k \in K} \in \mathbb{R}^m$, where $v_k > 0$ (resp. $v_k < 0$) denotes votes in favor of (resp. against) issue k. Casting v_k votes on issue k costs cv_k^2 voice credits, so that the total cost is

$$C(v) = c \sum_{k \in K} v_k^2,$$

with $c > 0$ a price parameter. The outcome on each issue is determined by the sign of aggregate votes, and voters choose v to maximize expected utility net of quadratic cost. In large populations, symmetric Bayesian equilibria implement approximately utilitarian allocations where the marginal price of influence on each issue is linear in v_k (Lalley and Weyl 2018; Georgescu et al. 2024).

5.3.1 A local quadratic-voting representation under fixed polarity

Return to our stylized Community Notes model. Fix n and treat the parameters $(\alpha, \beta, \mu, \theta)$ of all *other* voters and all notes as given. Suppose that the latent polarity θ_{i^*} of the focal voter is exogenously fixed and not re-estimated from i^*'s current ratings. In this counterfactual, the only part of the estimation problem affected by i^*'s ratings is the note-side intercept and loading parameters (μ, β), and the effect is mediated through the objective Q_n in (1).

Let $r = (r_j)_{j \in J} \in \{-1, 0, 1\}^m$ denote i^*'s rating vector, and let $\varphi(r)$ denote the resulting note-side parameters when Q_n is minimized holding θ_{i^*} fixed. Suppose that the focal voter has additively separable cardinal preferences u_j over the helpfulness scores of individual notes and that her utility is

$$U(r) = \sum_{j \in J} u_j\big(h_j(\varphi(r))\big) = \sum_{j \in J} u_j\big(\mu_j(r)\big).$$

We study small deviations from a baseline rating vector r^0. Because the penalized objective Q_n is smooth and strongly convex in φ and θ_{i^*} is fixed, the mapping $r \mapsto \varphi(r)$ is differentiable in a neighborhood of r^0 by the implicit function theorem. Thus $U(r)$ is twice continuously differentiable in a neighborhood of r^0 when we relax r to lie in $[-1, 1]^m$ and restrict attention to sufficiently small perturbations.

Lemma 5.5 (Local quadratic-voting form under fixed polarity)**.** *Fix the ratings of all voters $i \neq i^*$ and fix θ_{i^*} exogenously. Let r^0 be a baseline rating vector for i^* and suppose $U(r)$ is twice continuously differentiable in a neighborhood of r^0. Then for all sufficiently small deviations Δr we have the second-order expansion*

$$U(r^0 + \Delta r) = U(r^0) + g^\top \Delta r - \frac{1}{2} \Delta r^\top H \, \Delta r + o\big(\|\Delta r\|_2^2\big),$$

where $g = \nabla U(r^0) \in \mathbb{R}^m$ and H is a symmetric positive semidefinite matrix determined by the curvature of Q_n and the utility functions $(u_j)_{j \in J}$. In particular, up to second order, the focal voter's local problem is equivalent to choosing "votes" Δr on issues j to maximize linear benefits $g_j \Delta r_j$ subject to a quadratic cost $\frac{1}{2} \Delta r^\top H \, \Delta r$.

Proof sketch. Differentiability of $r \mapsto \varphi(r)$ follows from the first-order condition $\nabla_\varphi Q_n(\varphi(r), r) = 0$ and the non-singularity of the Hessian $\nabla^2_{\varphi\varphi} Q_n$ implied by strong convexity. The chain rule gives

$$\nabla U(r) = J(r)^\top \nabla_\varphi \tilde{U}(\varphi(r)),$$

where $\tilde{U}(\varphi) = \sum_j u_j(\mu_j)$ and $J(r) = \partial \varphi(r) / \partial r$ is the Jacobian. Differentiating once more yields the

Hessian

$$\nabla^2 U(r) = J(r)^\top \nabla^2_{\varphi\varphi} \tilde{U}(\varphi(r)) J(r) \; + \; \sum_j \big[\nabla_\varphi \tilde{U}(\varphi(r))\big]_j \frac{\partial^2 \varphi_j(r)}{\partial r \, \partial r^\top}.$$

At the baseline r^0, the first term is positive semidefinite because $\nabla^2_{\varphi\varphi} Q_n$ is positive definite while u_j is concave or locally approximated by its second-order Taylor expansion. The second term can be absorbed into the definition of H for sufficiently small perturbations. The Taylor expansion of U around r^0 then yields the stated representation. Full details, including an explicit expression for H in terms of the Hessian of Q_n, are provided in the appendix. □

Lemma 5.5 shows that, when θ_{i^*} is held fixed, the focal voter's local manipulation problem has the same structure as quadratic voting: she chooses a vector of marginal deviations Δr that yield linear benefits but are disciplined by a quadratic cost matrix H summarizing how her ratings must "fight against" the curvature of the global estimation problem. In this sense, her pattern of helpfulness ratings can be interpreted as an allocation of a finite "reputational currency" across issues, with a roughly quadratic tradeoff across notes.

5.4 Endogenous polarity and breakdown of the quadratic-voting analogy

In the actual Community Notes algorithm, the latent polarity θ_{i^*} of voter i^* is estimated jointly with all other parameters.(**CommunityNotesGuide**; B. Slaughter et al. 2025) Ratings that deviate from what the latent model predicts move θ_{i^*}, which in turn changes the predicted ratings for *all* notes. This feedback breaks the simple quadratic-voting analogy.

Formally, when θ_{i^*} is endogenous we can write the estimator as a mapping

$$r \mapsto \psi(r) = \big(\mu(r), \alpha(r), \beta(r), \theta(r)\big),$$

and the focal voter's utility as $U(r) = \sum_j u_j\big(\mu_j(r)\big)$. The Hessian of U with respect to r now contains additional terms involving $\partial\theta_{i^*}(r)/\partial r$ and $\partial^2\theta_{i^*}(r)/\partial r^2$. Even for small perturbations, the effective quadratic form need not be constant across directions in r and can depend on the baseline r^0 in a highly nonlinear way.

We illustrate the resulting nonlinearity with a simple fully-worked example based on a rank-one factorization of a 2×2 rating matrix.

Example 5.6 (Two-voter, two-note failure of quadratic voting). Consider two voters $i \in \{1, 2\}$ and two notes $j \in \{1, 2\}$. Let Y denote the 2×2 rating matrix with entries $y_{ij} \in \{0, 1\}$. Suppose the platform fits the unregularized rank-one model

$$y_{ij} \approx \beta_j \theta_i$$

by solving

$$\min_{\theta,\beta} \sum_{i=1}^{2} \sum_{j=1}^{2} \big(y_{ij} - \beta_j \theta_i\big)^2 \quad \text{subject to} \quad \|\theta\|_2 = 1, \; \|\beta\|_2 = 1,$$

so that (θ, β) is the leading left and right singular vectors of Y.(Amatriain et al. 2011) Take the helpfulness score of note j to be $h_j = \beta_j$ and classify a note as selected whenever h_j exceeds a fixed threshold.

Fix the second voter's ratings as $(y_{21}, y_{22}) = (1, 1)$ and let the focal voter $i^* = 1$'s ratings be $(r_1, r_2) \in \{0, 1\}^2$. Thus the data matrix is

$$Y(r_1, r_2) = \begin{pmatrix} r_1 & r_2 \\ 1 & 1 \end{pmatrix}.$$

For each of the four possible rating vectors (r_1, r_2), compute the leading right singular vector of $Y(r_1, r_2)$ and scale it by the largest singular value to obtain helpfulness scores (h_1, h_2).[11] One obtains:

$$\begin{aligned}
(r_1, r_2) = (0, 0) \quad &\Rightarrow \quad (h_1, h_2) \approx (-1.00, -1.00), \\
(r_1, r_2) = (1, 0) \quad &\Rightarrow \quad (h_1, h_2) \approx (-1.376, -0.851), \\
(r_1, r_2) = (0, 1) \quad &\Rightarrow \quad (h_1, h_2) \approx (-0.851, -1.376), \\
(r_1, r_2) = (1, 1) \quad &\Rightarrow \quad (h_1, h_2) \approx (-1.414, -1.414).
\end{aligned}$$

Focus on the marginal effect of changing r_1 while holding r_2 fixed. When $r_2 = 0$, changing r_1 from 0 to 1 changes the note scores by

$$\Delta h^{(r_2=0)} = (-1.376, -0.851) - (-1.00, -1.00) \approx (-0.376, +0.149).$$

In contrast, when $r_2 = 1$, changing r_1 from 0 to 1 changes the scores by

$$\Delta h^{(r_2=1)} = (-1.414, -1.414) - (-0.851, -1.376) \approx (-0.563, -0.038).$$

Thus the marginal effect of a positive rating on note 1 depends strongly on whether the voter has also rated note 2 positively: for the second note, the sign of the spillover flips from positive ($+0.149$) to negative (-0.038) depending on r_2. This kind of cross-issue interaction cannot be captured by any model in which the voter buys independent votes on each issue at a fixed quadratic price.(Casella and Sanchez 2018)

Intuitively, the nonlinearity arises because the same change in r_1 shifts the estimated polarity θ_{i^*} by different amounts depending on the existing value of r_2, which in turn affects the fitted polarity loadings β_1, β_2 and hence both helpfulness scores. The focal voter's "marginal price" of influence on issue 1 is not linear in the number of positive ratings she has already cast, and it is not separable across issues.

Example 5.6 is deliberately simple, but it illustrates a general point: once θ_{i^*} is estimated endogenously, the mapping from the focal voter's ratings to note scores is highly nonlinear, with

[11] All computations in this example are straightforward linear algebra and are provided in full in the appendix.

cross-issue complementarities and substitutabilities generated by the joint factorization. The smooth local quadratic representation in Lemma 5.5 may still hold for very small perturbations, but the effective cost matrix H becomes state-dependent and the global structure of the problem no longer resembles quadratic voting.

5.5 Reputational currency and the difficulty of optimal behavior

The preceding analysis suggests an interpretation of Community Notes in terms of "reputational currency". The latent polarity θ_i and intercept α_i summarize how well voter i's past ratings fit into the low-rank latent structure inferred from the population. Ratings that are consistent with the latent model keep θ_i near the center, preserving the rater's future influence on many notes; ratings that are systematically off-model push θ_i towards the extremes, reducing the effective weight of the rater's votes in future note rankings.(**CommunityNotesGuide**; Warden 2024d) In this sense, raters "spend" reputational currency when they cast ratings that are informative in the latent model, and ratings that appear idiosyncratic or partisan may deplete this currency.

From the perspective of a sophisticated, optimizing voter i^* with cardinal preferences over which notes are selected, the best-response problem is to choose a rating vector r to maximize expected utility over the selected set:

$$\max_{r\in\{-1,0,1\}^m} U_{i^*}(r) \quad \text{where} \quad U_{i^*}(r) = u_{i^*}\big(S(\psi(r))\big)$$

and $S(\psi(r))$ is the selected set when the platform estimates ψ from all voters' ratings, including r. In general this is a highly opaque bi-level optimization problem:(Hardt et al. 2016; Sundaram et al. 2023)

- the inner problem is the global parameter estimation problem that solves (1);
- the outer problem maps the resulting parameters to a discrete set $S(\psi(r))$ and then to utility $U_{i^*}(r)$.

Even ignoring integer constraints on r and relaxing to $r \in [-1,1]^m$, $U_{i^*}(r)$ is typically non-convex and does not admit closed-form expression. Moreover, related optimization problems in strategic classification and incentive-aware recommender systems are known to be computationally hard.(Hardt et al. 2016; Jin et al. 2022; Sundaram et al. 2023; Dai 2024)

We formalize this difficulty in a conservative way by embedding a classical NP-hard problem into the best-response problem of a focal voter, in a stylized Community Notes environment.

Proposition 5.7 (Computational hardness of best responses)**.** *There exists a family of stylized Community Notes estimation problems of the form (1), with fixed note set J and fixed ratings of voters $i \neq i^*$, and a family of utility functions u_{i^*} such that the following decision problem is NP-hard:*

Instance: *An integer B and a description of u_{i^*} and of the other voters' ratings.*

Question: *Does there exist a rating vector* $r \in \{-1, 0, 1\}^m$ *for* i^* *with at most* B *non-zero entries such that* $U_{i^*}(r) \geq \bar{U}$ *for a given threshold* $\bar{U}$*?*

Proof sketch. We reduce from a standard NP-hard problem such as Max-2-Sat or Max-Cut. Each decision variable in the NP-hard problem is mapped to a note, and each clause (or edge) is mapped to a pattern of ratings by the non-focal voters. The platform's estimator is constructed so that, for any rating vector r of the focal voter, the selected set $S(\psi(r))$ encodes the set of satisfied clauses under the implied truth assignment, with $U_{i^*}(r)$ proportional to the number of satisfied clauses.

Concretely, one can build an instance of (1) whose minimizer coincides with the solution of a linear separator in a strategic classification problem of the type studied by Hardt et al. (2016) and Sundaram et al. (2023). The latter show that computing an optimal incentive-aware empirical risk minimizer is NP-hard under natural assumptions on preferences and costs.(Sundaram et al. 2023) By embedding their SERM instance into the Community Notes parameterization, any polynomial-time algorithm that solved the best-response problem in our environment would solve an NP-hard problem. Full details of the reduction, including the mapping from clauses to voters and notes, are relegated to the appendix.

The restriction to at most B non-zero ratings enforces a budget constraint analogous to a limit on the number of issues on which i^* can intervene, and does not affect NP-hardness. □

Proposition 5.7 should be interpreted qualitatively rather than as a literal claim about the deployed Community Notes system. It highlights that, even in a stylized version of the algorithm with a single latent factor and quadratic regularization, computing an optimal rating strategy for a single voter can be as hard as solving a canonical NP-hard combinatorial problem. Combined with the nonlinearity illustrated in Example 5.6, this suggests that real-world raters cannot be expected to compute (or even approximate) their best responses in any systematic way. Instead, they behave myopically or heuristically, while the platform's algorithm effectively converts their past rating history into a stock of reputational currency whose marginal value is hard to foresee.(Dai 2024; Björkegren 2024)

6 Singling Out One Issue

This section focuses on the marginal effect of a single note on the Community Notes algorithm. We fix the set of voters and treat all other issues as a background environment. This allows us to formalize the influence of one column of the rating matrix on the estimated latent parameters and on which other notes are surfaced.

6.1 Set-up: fixing voters and a focal note

Let $I = \{1, \ldots, n\}$ denote the finite set of voters (contributors) and, for each $k \geq 1$, let $M^{(k)} = \{1, \ldots, k\}$ denote the set of candidate notes (issues). For expositional convenience we take the k-th

note to be the focal note and write $m^* = k$. For each k we observe a (possibly sparse) rating matrix

$$R^{(k)} = (r_{im})_{i \in I,\, m \in M^{(k)}} \in [-1, 1]^{n \times k},$$

where r_{im} is voter i's rating of note m and missing entries correspond to non-participation.[12]

As in the stylized Community Notes model introduced earlier, we assume an approximate factor structure for the rating matrix. In its simplest linear form, this takes the shape

$$R^{(k)} = \mathbf{1}_n \mu^{(k)\top} + U\, V^{(k)\top} + E^{(k)}, \tag{2}$$

where

- $U \in \mathbb{R}^{n \times d}$ collects latent voter factors $u_i^\top$ as its rows,
- $V^{(k)} \in \mathbb{R}^{k \times d}$ collects latent note factors $v_m^\top$ as its rows,
- $\mu^{(k)} \in \mathbb{R}^k$ is a note-specific intercept vector,
- $E^{(k)}$ is an idiosyncratic error matrix with mean zero and bounded moments.

The factor structure in (2) should be viewed as an approximation to the matrix-factorization model implemented by the Community Notes algorithm, which is estimated via regularized logistic matrix factorization on a note–rater helpfulness matrix rather than linear least squares.[13]

We collect all latent parameters in the vector

$$\theta^{(k)} = \big(U, V^{(k)}, \mu^{(k)}\big).$$

The Community Notes estimation step is represented abstractly by a map

$$T_k : \mathcal{R}_k \to \Theta_k, \qquad \widehat{\theta}^{(k)} = T_k\big(R^{(k)}\big), \tag{3}$$

where $\mathcal{R}_k$ is the set of feasible rating matrices of size $n \times k$ and Θ_k is the parameter space. In the concrete implementation, T_k is the output of a regularized maximum-likelihood routine for a logistic factor model; in this section we only require that T_k satisfy the regularity conditions stated below.

Given estimated parameters $\widehat{\theta}^{(k)}$, Community Notes computes a *bridging score* for each note m, designed to reward notes that are found helpful by voters with diverse viewpoints.[14] We model this

[12] The arguments below can be extended to fully general missingness patterns; for simplicity we treat $R^{(k)}$ as an $n \times k$ matrix with bounded entries.

[13] See Wojcik, Hilgard, Judd, Mocanu, Ragain, Hunzaker, et al. (2022a) for the original Birdwatch implementation and Community Notes Team (2023) for the current open-source Community Notes scorer, both of which rely on a low-rank factor representation of the note–rater matrix.

[14] The current production system constructs predicted helpfulness probabilities for each note and for each voter using the estimated latent factors, and then aggregates these predictions separately for different latent viewpoint groups before applying a minimum-type scoring rule; see Wojcik, Hilgard, Judd, Mocanu, Ragain, Hunzaker, et al. (2022a) and Community Notes Team (2023).

step as a deterministic scoring function

$$s_m^{(k)} = s_m^{(k)}\big(R^{(k)}, \widehat{\theta}^{(k)}\big) \in \mathbb{R}, \qquad m \in M^{(k)}, \tag{4}$$

and a *selection rule* that determines which notes are surfaced:

$$S_k\big(R^{(k)}\big) \;=\; \big\{m \in M^{(k)} : s_m^{(k)} \geq \tau_k \text{ and } m \text{ passes the guardrails}\big\}, \tag{5}$$

where τ_k is a (possibly k-dependent) score threshold and the guardrails encode requirements such as a minimum number of ratings and checks for policy-violating content.

For a fixed k, we now *freeze* the rating behavior of all voters on all notes other than m^*. Formally, write

$$R^{(k)} = \big(R^{(-m^*)}, r_{\cdot m^*}\big),$$

where $R^{(-m^*)}$ collects all columns other than m^* and $r_{\cdot m^*}$ is the m^*-th column. We will study how changes to $r_{\cdot m^*}$—or the removal of the column m^*—affect the estimated parameters $T_k(R^{(k)})$ and the surfaced set $S_k(R^{(k)})$.

6.2 Vanishing marginal influence of one issue for large k

We now show that, under standard large-k regularity conditions for approximate factor models, the marginal influence of a single note on the estimated parameters and on the selection of *other* notes vanishes as the number of notes grows.

To make this precise, fix n and consider two rating matrices $R^{(k)}, \widetilde{R}^{(k)} \in \mathcal{R}_k$ that agree on all but the focal column:

$$R^{(k)} = \big(R^{(-m^*)}, r_{\cdot m^*}\big), \qquad \widetilde{R}^{(k)} = \big(R^{(-m^*)}, \tilde{r}_{\cdot m^*}\big).$$

Let

$$\widehat{\theta}^{(k)} = T_k\big(R^{(k)}\big), \qquad \widetilde{\theta}^{(k)} = T_k\big(\widetilde{R}^{(k)}\big),$$

and similarly denote by $s_m^{(k)}$ and $\tilde{s}_m^{(k)}$ the resulting scores.

We impose the following regularity conditions, which are satisfied by principal-components estimators for linear approximate factor models and, more generally, by a large class of maximum-likelihood factor estimators.[15]

Assumption 3.1 (Approximate factor model). For each k, the columns $\{r_{\cdot m}\}_{m \in M^{(k)}}$ are independent draws from a common distribution with $\mathbb{E}[r_{\cdot m}] = \mu^\star$, bounded entries $\|r_{\cdot m}\|_\infty \leq R_{\max}$ almost surely, and covariance matrix $\Sigma_r = \Lambda\Lambda^\top + \Psi$, where $\Lambda \in \mathbb{R}^{n \times d}$ has full column rank, d is fixed, and Ψ is diagonal with bounded entries.

[15] See, among many others, Stock and Watson (2002), Bai (2003a), and Yu et al. (2015) for asymptotic theory for principal-components factor estimation and eigenvector perturbation bounds that translate small changes in the empirical second-moment matrix into small changes in the estimated factor space.

Assumption 3.2 (Spectral gap). The nonzero eigenvalues of Σ_r satisfy $\lambda_1(\Sigma_r) \geq \cdots \geq \lambda_d(\Sigma_r) > \lambda_{d+1}(\Sigma_r)$ and the eigen-gap $\delta = \lambda_d(\Sigma_r) - \lambda_{d+1}(\Sigma_r)$ is strictly positive.

Assumption 3.3 (Lipschitz dependence on empirical moments). There exists a function Φ_k such that $\widehat{\theta}^{(k)} = \Phi_k(\widehat{\Sigma}^{(k)})$, where

$$\widehat{\Sigma}^{(k)} = \frac{1}{k}\sum_{m=1}^{k}(r_{\cdot m} - \overline{r}^{(k)})(r_{\cdot m} - \overline{r}^{(k)})^\top, \qquad \overline{r}^{(k)} = \frac{1}{k}\sum_{m=1}^{k} r_{\cdot m},$$

and Φ_k is uniformly Lipschitz in operator norm:

$$\big\|\Phi_k(A) - \Phi_k(B)\big\| \leq L\,\|A - B\|_{\mathrm{op}} \quad \text{for all symmetric } A, B,$$

for some constant L independent of k.

Assumption 3.4 (Smooth scoring rule). For each m the score function can be written as $s_m^{(k)} = g_m\big(\widehat{\theta}^{(k)}\big)$ for some function $g_m : \Theta_k \to \mathbb{R}$ that is Lipschitz with constant at most L_s, uniformly in k and m.

Assumption 3.3 abstracts the fact that principal-components and related factor estimators depend on the data only through the empirical covariance (and possibly empirical means), and that eigenvector perturbations are Lipschitz in the perturbation of the covariance matrix when there is an eigen-gap.[16] Assumption 3.4 matches the implementation where the scoring rule is a smooth functional of the estimated factors, such as averages of predicted helpfulness across latent groups.

We measure the difference between two parameter vectors by a norm $\|\cdot\|_{\Theta_k}$ on Θ_k. For concreteness, the reader may take $\|\cdot\|_{\Theta_k}$ to be the Frobenius norm on the stacked matrices $(U, V^{(k)}, \mu^{(k)})$, modulo the usual rotational invariance of the factor space.

Proposition 6.1 (One-issue influence bound)**.** *Suppose Assumptions 3.1–3.4 hold and $R_{\max} < \infty$. Then there exists a constant $C < \infty$, independent of k, such that for all $k \geq 2$,*

$$\big\|\widehat{\theta}^{(k)} - \widetilde{\theta}^{(k)}\big\|_{\Theta_k} \;\leq\; \frac{C}{k}, \tag{6}$$

whenever $R^{(k)}$ and $\widetilde{R}^{(k)}$ differ in at most one column. Moreover, for every note $m \neq m^*$,

$$\big|s_m^{(k)} - \tilde{s}_m^{(k)}\big| \;\leq\; \frac{CL_s}{k}. \tag{7}$$

Consequently, for any fixed note $m \neq m^*$ and any threshold sequence τ_k that stays away from the score of m in the limit,

$$\mathbb{P}\Big(\mathbf{1}\{m \in S_k(R^{(k)})\} = \mathbf{1}\{m \in S_k(\widetilde{R}^{(k)})\}\Big) \longrightarrow 1 \qquad \text{as } k \to \infty.$$

[16] A precise statement for eigenvectors is provided, for instance, by the variant of the Davis–Kahan theorem in Yu et al. (2015).

Proof sketch. Changing a single column of the rating matrix affects the empirical covariance $\widehat{\Sigma}^{(k)}$ by at most a rank-two perturbation of size $O(1/k)$ in operator norm. Indeed, since $\|r_{\cdot m}\|_2 \le \sqrt{n} R_{\max}$ for all m by boundedness, one has

$$\big\|\widehat{\Sigma}^{(k)} - \widetilde{\Sigma}^{(k)}\big\|_{\mathrm{op}} \;\le\; \frac{2nR_{\max}^2}{k},$$

where $\widehat{\Sigma}^{(k)}$ and $\widetilde{\Sigma}^{(k)}$ denote the empirical covariance matrices computed from $R^{(k)}$ and $\widetilde{R}^{(k)}$, respectively.

By Assumption 3.3, this implies

$$\big\|\widehat{\theta}^{(k)} - \widetilde{\theta}^{(k)}\big\|_{\Theta_k} \;\le\; L\,\big\|\widehat{\Sigma}^{(k)} - \widetilde{\Sigma}^{(k)}\big\|_{\mathrm{op}} \;\le\; \frac{C}{k}$$

for some C that depends only on n, $R_{\max}$ and L, yielding (6). The bound (7) follows directly from the Lipschitz property of the scoring functions in Assumption 3.4.

For the selection indicators, fix $m \neq m^*$ and suppose that its limiting score s_m^∞ exists and is such that $|s_m^\infty - \tau_k| \ge \varepsilon$ for some $\varepsilon > 0$ and all large k. Combined with (7), this implies that $s_m^{(k)}$ and $\tilde{s}_m^{(k)}$ eventually lie on the same side of τ_k with probability tending to one, so the selection indicator for m is eventually the same under both $R^{(k)}$ and $\widetilde{R}^{(k)}$. A more detailed argument, relegated to the appendix, shows that a similar conclusion holds uniformly over all $m \neq m^*$ except for a vanishing fraction of near-threshold notes. □

Proposition 6.1 formalizes the intuition that, in a large factor model, the contribution of any single column to the empirical second-moment matrix—and hence to the estimated latent factors and note parameters—is of order $1/k$. The presence or rating pattern of a single note therefore has a vanishing marginal influence on the estimated parameters and on the fate of other notes, except in knife-edge cases where another note's score lies arbitrarily close to the surfacing threshold.

6.3 Community Notes as computing a counterfactual center

The factor representation in (2) can be used to interpret Community Notes as estimating and applying a form of *counterfactual center* of the approval distribution. To make this concrete, it is helpful to work with a one-dimensional latent viewpoint coordinate.

Assume that the first column of U encodes an ideological or viewpoint coordinate $u_i \in \mathbb{R}$ for each voter i, normalized so that $\mathbb{E}[u_i \mid i \text{ participates}] = 0$ and $\mathrm{Var}(u_i \mid i \text{ participates}) = 1$. Write u_i for this scalar coordinate and let the remaining columns of U capture residual heterogeneity. For notational simplicity we focus on the dependence of ratings on u_i and suppress other factors.

For each note m, consider the *latent helpfulness profile*

$$h_m(u) = \mathbb{E}\big[r_{im} \mid u_i = u\big],$$

defined for all u in the support of the participation-weighted distribution of viewpoints. This

function summarizes how voters at viewpoint u would evaluate note m in expectation. In the logistic matrix-factorization implementation, $h_m(u)$ is approximated by a link function applied to an affine function of u, induced by the note- and voter-specific latent factors estimated from the rating matrix (Wojcik, Hilgard, Judd, Mocanu, Ragain, Hunzaker, et al. 2022a; Community Notes Team 2023).

We now define a notion of center that is intrinsic to the population of participating voters.

Definition 6.2 (Counterfactual center of participation)**.** Let F denote the distribution of the viewpoint coordinate u_i among voters who participate in Community Notes. Fix an even, strictly convex loss function $\ell : \mathbb{R} \to \mathbb{R}_+$ with $\ell(0) = 0$ and define

$$c^* \in \arg\min_{c \in \mathbb{R}} \int \ell(u - c)\, \mathrm{d}F(u).$$

Any minimizer c^* is called a *counterfactual center of participation*. When $\ell(x) = x^2$ this is the participation-weighted mean; when $\ell(x) = |x|$ it is a generalized median.

The terminology "counterfactual" reflects the fact that c^* need not correspond to any particular individual voter; it is a hypothetical viewpoint summarizing the distribution of those who choose to participate. In practice, the algorithm does not compute c^* as a separate object; rather, it learns latent coordinates u_i and then aggregates predicted helpfulness in a way that is approximately equivalent to evaluating notes from the perspective of this center.

Given c^*, an idealized *central approval score* for note m is

$$a_m^{\mathrm{ctr}} = \int h_m(u)\, w(u - c^*)\, \mathrm{d}F(u), \tag{8}$$

where w is a nonnegative, symmetric weighting function with $w(0) = 1$ and $\int w(u - c^*)\, \mathrm{d}F(u) = 1$. The choice $w(u - c^*) \propto \mathbf{1}\{|u - c^*| \le \delta\}$ corresponds to averaging over a band of "moderate" voters, while a smoother w downweights extreme viewpoints more gradually.

In the idealized setting where the full latent helpfulness profiles $\{h_m\}_m$ are known, a natural *central selection rule* would surface those notes whose central approval exceeds a threshold:

$$S^{\mathrm{ctr}} = \{m : a_m^{\mathrm{ctr}} \ge \tau\}. \tag{9}$$

The Community Notes algorithm can be interpreted as an implementable approximation to (9). The factor model $R^{(k)} \approx UV^{(k)\top}$ provides point estimates $\widehat{u}_i$ of the viewpoint coordinate and predicted helpfulness profiles $\widehat{h}_m(u)$. The current bridging-based scoring rule, as implemented in the open-source scorer, aggregates predicted helpfulness separately for voters on opposite sides of the latent axis and then applies a minimum-type functional to favor notes that are broadly accepted across that axis.[17]

[17] In particular, the production scorer uses the estimated latent factors to predict helpfulness probabilities for all voters, averages these predictions within latent groups that correspond roughly to "left" and "right" halves of the distribution of $\widehat{u}_i$, and then constructs a "bridging" score that increases when both group averages are high (Wojcik, Hilgard, Judd, Mocanu, Ragain, Hunzaker, et al. 2022a; Community Notes Team 2023; Ovadya and Thorburn 2023).

To see the connection, suppose that F is symmetric around c^* and that each h_m is smooth with bounded first derivative. Define the idealized bridging score

$$b_m = \min\Big\{\mathbb{E}\big[h_m(u) \mid u \le c^*\big]\,\mathbb{E}\big[h_m(u) \mid u \ge c^*\big\}\Big\}.$$

For small departures from symmetry and for notes whose helpfulness profiles are not strongly polarized, b_m and a_m^{ctr} rank notes in approximately the same way:

Proposition 6.3 (Bridging score as a central approval functional)**.** *Suppose F is symmetric around c^* and h_m is continuously differentiable with Lipschitz derivative uniformly over m. Then for any two notes m and $\tilde{m}$ satisfying $h_m(u) \ge h_{\tilde{m}}(u)$ for all u in a neighborhood of c^* and $h_m(u) \ge h_{\tilde{m}}(u)$ for all u with $|u - c^*|$ sufficiently large, one has*

$$a_m^{\mathrm{ctr}} \ge a_{\tilde{m}}^{\mathrm{ctr}} \quad \Longrightarrow \quad b_m \ge b_{\tilde{m}}.$$

Moreover, whenever h_m and $h_{\tilde{m}}$ differ strictly on a set of positive F-measure near c^*, the implication is strict: $a_m^{\mathrm{ctr}} > a_{\tilde{m}}^{\mathrm{ctr}}$ implies $b_m > b_{\tilde{m}}$.

Proof sketch. By symmetry of F around c^* and the Lipschitz properties of h_m, the difference $a_m^{\mathrm{ctr}} - a_{\tilde{m}}^{\mathrm{ctr}}$ can be written as a weighted average of $h_m(u) - h_{\tilde{m}}(u)$ over the left and right halves of the distribution, with weights that are themselves symmetric. The assumption that h_m dominates $h_{\tilde{m}}$ both near c^* and in the tails implies that this average is nonnegative. The bridging score b_m is the minimum of the left and right averages, so $b_m \ge b_{\tilde{m}}$ whenever both left and right averages for m are at least as large as those for $\tilde{m}$. The strict inequality follows when $h_m - h_{\tilde{m}}$ is strictly positive on a set of positive measure near c^*. A formal argument, which again uses smoothness and symmetry, is provided in the appendix. □

Proposition 6.3 says that, under mild shape restrictions on helpfulness profiles, the bridging-based scoring rule implemented by Community Notes selects notes in a way that is monotone in their central approval a_m^{ctr}. Equivalently, the algorithm can be viewed as estimating a latent viewpoint axis and then approximating the hypothetical decision rule that would surface notes selected by a center-of-participation aggregator.

6.4 Why disagreement in the past need not be reconciled

A common informal slogan about Community Notes is that *people who have disagreed in the past need to agree on a note for it to be surfaced.* This slogan captures the spirit of rewarding cross-cutting agreement, but it is not literally true, even at the level of individual issues. The algorithm uses historical disagreement to estimate a latent axis and viewpoint distribution, and then it privileges notes that attract agreement from *centrally located* users on both sides of that axis. It does not require that all historically opposed individuals agree on each surfaced note.

We make this precise by considering two voters who frequently disagreed in the past and showing that, for a given note, they can be dominated by a sufficiently large group of centrists.

Consider three disjoint sets of voters:

$$L = \{i : u_i \le -a\}, \qquad C = \{i : |u_i| < a\}, \qquad R = \{i : u_i \ge a\},$$

where $a > 0$ and u_i is the latent viewpoint coordinate. Suppose L and R together contain two extreme voters $i_L \in L$ and $i_R \in R$ whose past ratings indicate persistent disagreement; for concreteness, assume that on a large fraction of notes m one has $\operatorname{sign}(r_{i_L m}) = -\operatorname{sign}(r_{i_R m})$. The factor model will therefore place i_L and i_R on opposite sides of the latent axis.

For a given note m, define the (idealized) group-level average helpfulness

$$\bar{h}_L(m) = \frac{1}{|L|} \sum_{i \in L} \widehat{h}_{im}, \quad \bar{h}_R(m) = \frac{1}{|R|} \sum_{i \in R} \widehat{h}_{im},$$

where $\widehat{h}_{im}$ is the predicted helpfulness probability for voter i on note m produced by the factor model. In the actual implementation these averages are taken over all voters with $\widehat{u}_i$ below or above a data-dependent threshold and use only a subset of voters who have rated enough notes; this does not affect the logic of the argument.

Proposition 6.4 (Centrists can dominate persistent disagreement)**.** *Fix $a > 0$ and assume that $|C|$ grows proportionally with the total number of voters, while the number of extreme voters in $L \cup R$ remains bounded. Suppose that for some note m:*

1. *All centrists find the note helpful: $\widehat{h}_{im} \approx 1$ for all $i \in C$.*
2. *The two extremists disagree: $\widehat{h}_{i_L m} \approx 1$ and $\widehat{h}_{i_R m} \approx 0$.*
3. *All other voters in $L \cup R$ have predicted helpfulness bounded away from zero.*

Then for any threshold $\tau \in (0, 1)$, there exists N such that whenever $|C| \ge N$,

$$\bar{h}_L(m) \ge \tau \quad \text{and} \quad \bar{h}_R(m) \ge \tau,$$

so that m is surfaced by a bridging-based rule that requires both group averages to exceed τ, even though i_L and i_R disagree on m.

Proof. Write

$$\bar{h}_L(m) = \frac{1}{|L|}\Big(\widehat{h}_{i_L m} + \sum_{i \in L \setminus \{i_L\}} \widehat{h}_{im}\Big),$$

and similarly for $\bar{h}_R(m)$. By assumption, all but finitely many voters in L and R have predicted helpfulness bounded below by some $\eta > 0$. Thus, for sufficiently large $|L|$ and $|R|$ the contribution of any fixed number of extreme voters is at most $O(1/|L|)$ or $O(1/|R|)$. In particular, the single disagreeing extremist i_R contributes at most $1/|R|$ to $\bar{h}_R(m)$, which can be made arbitrarily small as $|R|$ grows.

On the other hand, assumption (1) implies that the large centrist group C, part of which is included in each side's aggregation in the production scorer, pushes both $\bar{h}_L(m)$ and $\bar{h}_R(m)$ arbitrarily close to one as $|C|$ grows. Hence for any fixed $\tau < 1$ there exists N such that whenever $|C| \geq N$ the group averages exceed τ, even though i_L and i_R continue to disagree on m. □

Proposition 6.4 shows that the strong slogan "people who have disagreed in the past must agree" fails as a description of the Community Notes selection rule viewed at the issue level. Historical disagreement is crucial for *learning* a latent axis and embedding users along that axis, but once this geometry is in place, the algorithm effectively privileges agreement among centrally located users on both sides of the divide. Extreme users who are far out on the latent axis, even if they have a long history of disagreement with one another, can be outweighed by a large group of centrist voters whose ratings dominate the group averages that feed into the bridging score.

From the perspective of bridging, this is exactly the intended behavior. The system uses patterns of past disagreement to infer a latent ideological dimension and then surfaces notes that provide common ground across moderate regions of that dimension, rather than notes that reconcile the most extreme viewpoints (Ovadya 2022b; Ovadya and Thorburn 2023). The "bridge" is thus built between central clusters on either side, not necessarily between the most polarized individuals.

7 Community Notes as a Statistical Mechanism

This section treats Community Notes as a statistical estimator of latent note quality and then embeds the estimator in a model of costly participation. Throughout, a "note" $m \in \{1, \ldots, M\}$ is a potential annotation and $i \in \{1, \ldots, I\}$ indexes users. The Community Notes algorithm is interpreted as a matrix-factorization estimator that maps the sparse rating matrix into predicted approval scores for notes, in the spirit of latent factor models used in recommender systems (Koren et al. 2009; Wojcik, Hilgard, Judd, Mocanu, Ragain, Hunzaker, et al. 2022b). Under correct specification and exogenous missingness, this estimator is asymptotically efficient in the usual sense of econometrics of factor models (Bai 2003b; Anatolyev and Mikusheva 2021). Once participation is endogenous, however, it converges to the average approval *among participants*, which can be systematically different from the population welfare benchmark familiar from costly voting models (Palfrey and Rosenthal 1985; Börgers 2004).

7.1 Estimation without selection bias

For each ordered pair (i, m), let $Y_{im} \in \mathbb{R}$ be user i's latent approval or "helpfulness" signal for note m. We interpret Y_{im} as an underlying quantitative rating (for example, an internal latent score that generates observed discrete ratings). The object of interest for note m is its population-average approval

$$\theta_m \equiv \mathbb{E}[Y_{im}],$$

where the expectation is taken over the population of users i.[18]

Let $D_{im} \in \{0, 1\}$ indicate whether user i rated note m. This subsection assumes *no selection bias*: conditional on the latent signal, missingness is independent of approval:

$$D_{im} \sim \text{i.i.d. Bernoulli}(\rho), \qquad D_{im} \perp Y_{im}. \tag{10}$$

Thus each observed rating for note m is an i.i.d. draw from the population of potential raters.

7.1.1 A simple random-effects factor structure

To compare Community Notes to a naive estimator, it is convenient to start from a one-factor random-effects structure capturing heterogeneous rater tendencies:

$$Y_{im} = \theta_m + a_i + \varepsilon_{im}, \tag{11}$$

where θ_m is the note-specific mean approval, a_i is a rater-specific effect (e.g. overall lenience, ideology, or generosity), and ε_{im} is an idiosyncratic error. Assume

$$\mathbb{E}[a_i] = 0, \quad \mathbb{V}(a_i) = \sigma_a^2, \qquad \mathbb{E}[\varepsilon_{im}] = 0, \quad \mathbb{V}(\varepsilon_{im}) = \sigma_\varepsilon^2, \tag{12}$$

with $\{a_i\}$ i.i.d., $\{\varepsilon_{im}\}$ i.i.d., and $(a_i, \varepsilon_{im}, D_{im})$ mutually independent across i and m. The rater effect a_i is a degenerate one-dimensional latent factor; richer matrix-factorization models simply replace a_i by low-dimensional vectors and introduce note-specific factor loadings (Koren et al. 2009).

For note m, let $S_m \equiv \sum_{i=1}^{I} D_{im}$ be the number of observed ratings and $R_m \equiv \{i : D_{im} = 1\}$ the set of raters. The naive estimator of θ_m is the sample mean of observed ratings:

$$\hat{\theta}_m^{\text{avg}} \equiv \frac{1}{S_m} \sum_{i \in R_m} Y_{im}. \tag{13}$$

Under (10)–(12) and $I \to \infty$ so that $S_m \to \infty$ in probability, $\hat{\theta}_m^{\text{avg}}$ is consistent and asymptotically normal with

$$\sqrt{S_m}\big(\hat{\theta}_m^{\text{avg}} - \theta_m\big) \xrightarrow{d} \mathcal{N}\big(0,\, \sigma_a^2 + \sigma_\varepsilon^2\big). \tag{14}$$

Intuitively, the variance of the naive estimator includes both between-rater heterogeneity σ_a^2 and idiosyncratic noise σ_ε^2.

Community Notes instead exploits the full rating matrix to estimate rater-specific latent factors and remove systematic heterogeneity across raters, as in the matrix-factorization literature (Koren et al. 2009; Wojcik, Hilgard, Judd, Mocanu, Ragain, Hunzaker, et al. 2022b). Formally, suppose each rater evaluates many notes, with M_i denoting the set of notes rated by user i and $|M_i| \to \infty$ as $M \to \infty$. Let $\hat{a}_i$ be (a component of) the estimated latent factor for rater i, obtained by (penalized)

[18] In practice, θ_m can be interpreted as the central tendency—for example, mean or median—of a bounded approval scale. The analysis below focuses on the mean for simplicity.

least squares or maximum likelihood on the full rating matrix.[19] Assume

$$\max_{1\le i\le I}\left|\hat{a}_i - a_i\right| = o_p\big(S_m^{-1/2}\big),$$

so the estimation error in $\hat{a}_i$ is negligible compared to the sampling variability in (14).

The Community Notes estimator for θ_m can then be idealized as the de-biased mean

$$\hat{\theta}_m^{\mathrm{CN}} \equiv \frac{1}{S_m}\sum_{i\in R_m}\big(Y_{im} - \hat{a}_i\big). \tag{15}$$

In more realistic factor models the subtraction uses a linear combination of estimated user factors rather than a single scalar $\hat{a}_i$, but the logic is the same: aggregate residualized ratings.

Proposition 7.1 (Efficiency gain from exploiting the factor structure)**.** *Under* (10)–(12) *and* (7.1.1)*, for fixed note m and $I \to \infty$:*

1. *Both estimators are consistent:*

$$\hat{\theta}_m^{\mathrm{avg}} \xrightarrow{p} \theta_m, \qquad \hat{\theta}_m^{\mathrm{CN}} \xrightarrow{p} \theta_m.$$

2. *Their asymptotic distributions satisfy*

$$\sqrt{S_m}\big(\hat{\theta}_m^{\mathrm{avg}} - \theta_m\big) \xrightarrow{d} \mathcal{N}\big(0,\, \sigma_a^2 + \sigma_\varepsilon^2\big), \tag{16}$$

$$\sqrt{S_m}\big(\hat{\theta}_m^{\mathrm{CN}} - \theta_m\big) \xrightarrow{d} \mathcal{N}\big(0,\, \sigma_\varepsilon^2\big). \tag{17}$$

In particular, if $\sigma_a^2 > 0$ then

$$\mathbb{V}_\infty\big(\hat{\theta}_m^{\mathrm{CN}}\big) < \mathbb{V}_\infty\big(\hat{\theta}_m^{\mathrm{avg}}\big),$$

so the Community Notes estimator is asymptotically more efficient than the naive estimator.

Proof. By the random-effects representation (11) and the fact that $\{i : D_{im} = 1\}$ is an i.i.d. subsample of users, we can write

$$\hat{\theta}_m^{\mathrm{avg}} - \theta_m = \frac{1}{S_m}\sum_{i\in R_m}(a_i + \varepsilon_{im}).$$

Because (a_i, ε_{im}) are i.i.d. with mean zero and finite second moments, a standard Lindeberg–Feller central limit theorem yields (16), using $\mathbb{V}(a_i + \varepsilon_{im}) = \sigma_a^2 + \sigma_\varepsilon^2$. Consistency follows from the weak law of large numbers.

[19] Under classical approximate factor-model conditions, principal-components or maximum-likelihood estimators of the latent factors are consistent and asymptotically normal as both the cross-sectional and note dimensions grow (Bai 2003b; Anatolyev and Mikusheva 2021).

For $\hat{\theta}_m^{\mathrm{CN}}$, add and subtract a_i:

$$\hat{\theta}_m^{\mathrm{CN}} - \theta_m = \frac{1}{S_m} \sum_{i \in R_m} (\varepsilon_{im}) + \frac{1}{S_m} \sum_{i \in R_m} (a_i - \hat{a}_i).$$

The first term is an average of i.i.d. mean-zero errors with variance σ_ε^2; the CLT gives the limit in (17). The second term is $o_p(S_m^{-1/2})$ by (7.1.1), so it is asymptotically negligible and does not affect the limiting distribution. Again, consistency follows by the law of large numbers. Because $\sigma_\varepsilon^2 < \sigma_a^2 + \sigma_\varepsilon^2$ whenever $\sigma_a^2 > 0$, the Community Notes estimator has strictly smaller asymptotic variance. □

Proposition 7.1 formalizes the intuitive gain from sharing statistical strength across notes and raters. In more general matrix-factorization models, the Community Notes estimator coincides with the best linear unbiased (or asymptotically efficient) estimator of the latent mean approval θ_m under the maintained factor structure (Bai 2003b; Anatolyev and Mikusheva 2021).

7.2 Estimation with endogenous participation

The previous subsection assumed missing ratings are independent of preferences. In practice, users choose which notes to rate. Extending the random-effects model, write again

$$Y_{im} = \theta_m + a_i + \varepsilon_{im},$$

but now allow the participation decision D_{im} to depend on user i's preferences or latent position. Formally, suppose

$$D_{im} = \mathbb{1}\{\gamma_m U_{im} - C_{im} + \eta_{im} \geq 0\}, \tag{18}$$

where U_{im} is user i's (cardinal) utility from note m being surfaced, C_{im} is the opportunity cost of rating, η_{im} is an idiosyncratic shock, and γ_m captures how strongly perceived utility translates into the propensity to rate. We allow U_{im} to be stochastically related to Y_{im}, so selection is *endogenous*.

Unlike the missing-completely-at-random assumption (10), (18) implies

$$\mathbb{E}[Y_{im} \mid D_{im} = 1] \neq \mathbb{E}[Y_{im}]$$

whenever U_{im} is correlated with Y_{im}. This is the canonical sample-selection setting familiar from labor-supply estimation (Heckman 1979). Neither the naive estimator nor the Community Notes estimator identifies the population mean θ_m without further structure on the selection process.

What is identified is the *average approval among participants*. Let

$$\theta_m^{\mathrm{part}} \equiv \mathbb{E}[Y_{im} \mid D_{im} = 1]. \tag{19}$$

Under (18), θ_m^{part} depends on the joint distribution of $(Y_{im}, U_{im}, C_{im}, \eta_{im})$ and on the participation threshold, and hence on the costs and perceived benefits of voting.

In the presence of selection, the Community Notes algorithm continues to fit a factor model to the *conditional* distribution of observed ratings. Formally, index the parameters of the factor model by a vector β and let $\mu_{im}(\beta)$ denote the model-implied conditional expectation of Y_{im} given i and m when a rating is observed. The estimator $\hat{\beta}$ solves a sample moment condition of the form

$$\frac{1}{\sum_{i,m} D_{im}} \sum_{i,m} D_{im}\big(Y_{im} - \mu_{im}(\hat{\beta})\big) \frac{\partial \mu_{im}(\hat{\beta})}{\partial \beta} = 0, \tag{20}$$

for example as a first-order condition of least squares or maximum likelihood on the observed entries of the rating matrix (Koren et al. 2009; Wojcik, Hilgard, Judd, Mocanu, Ragain, Hunzaker, et al. 2022b). The population counterpart of (20) is

$$\mathbb{E}\Big[D_{im}\big(Y_{im} - \mu_{im}(\beta^*)\big) \frac{\partial \mu_{im}(\beta^*)}{\partial \beta}\Big] = 0, \tag{21}$$

which defines a pseudo-true parameter β^*.

Proposition 7.2 (Community Notes as an estimator of approval conditional on participation)**.** *Suppose:*

1. *The participation rule is given by (18) and induces strictly positive participation probabilities,* $0 < \Pr(D_{im} = 1) < 1$.
2. *For some* β^**,*
$$\mu_{im}(\beta^*) = \mathbb{E}[Y_{im} \mid i, m, D_{im} = 1],$$
so the factor model is correctly specified for the conditional mean among participants.
3. *Regularity conditions ensure that the solution* $\hat{\beta}$ *to (20) is consistent for* β^* *as the number of observed ratings grows.*

Then for each fixed note m*, the Community Notes prediction*

$$\hat{\theta}_m^{\mathrm{CN}} \equiv \hat{\theta}_m(\hat{\beta})$$

converges in probability to

$$\theta_m^{\mathrm{part}} = \mathbb{E}[Y_{im} \mid D_{im} = 1],$$

the average approval among participants. In general, $\theta_m^{\mathrm{part}} \neq \theta_m$*, and* θ_m *is not point-identified from Community Notes data without further assumptions on selection.*

Proof. Under (ii) and the definition of β^*, the population moment condition (21) holds:

$$\mathbb{E}\Big[D_{im}\big(Y_{im} - \mathbb{E}[Y_{im} \mid i, m, D_{im} = 1]\big) \frac{\partial \mu_{im}(\beta^*)}{\partial \beta}\Big] = 0.$$

By (iii), the estimator $\hat{\beta}$ converges in probability to β^*. For any fixed note m, the Community Notes prediction is a smooth function of $\hat{\beta}$, so

$$\hat{\theta}_m^{\mathrm{CN}} = \hat{\theta}_m(\hat{\beta}) \xrightarrow{p} \theta_m(\beta^*) = \mathbb{E}[Y_{im} \mid D_{im} = 1],$$

where the last equality uses (ii) and the definition (19). Unless D_{im} is independent of Y_{im} (the special case of missing completely at random), this conditional mean differs from the population mean θ_m.[20] □

Proposition 7.2 highlights that once participation is endogenous, Community Notes recovers the distribution of opinions of those who choose to vote, not the distribution in the underlying population. Whether this is normatively desirable depends on how participation costs relate to utilities and how strongly society wishes to weigh the preferences of non-participants.

7.3 Welfare with costly voting and cardinal preferences

We now introduce social welfare and analyze how Community Notes performs when rating is costly. Let there be a continuum of users indexed by type $x \in X$, with total mass one and distribution F on X. For a given note m, user x derives cardinal utility $u_m(x)$ if the note is surfaced (displayed) and zero otherwise. Let $S_m \in \{0, 1\}$ denote whether the note is ultimately surfaced by the platform's mechanism.

Each user may choose to rate note m or not. Let $d_m(x) \in \{0, 1\}$ denote the participation decision of type x, with $d_m(x) = 1$ indicating that x contributes a rating and pays a cost $c_m(x) \geq 0$. Ratings are mapped into a mechanism-specific statistic, such as the Community Notes latent approval estimate $\hat{\theta}_m^{\mathrm{CN}}$, and S_m is determined by a threshold rule:

$$S_m = \mathbb{1}\{\hat{\theta}_m \geq \tau\} \tag{22}$$

for some cutoff τ, where $\hat{\theta}_m$ is either $\hat{\theta}_m^{\mathrm{CN}}$ or the simple approval mean defined below.

7.3.1 Social welfare

Fix a note m and suppress the index m when no confusion arises. Given a participation profile $d(\cdot)$ and induced surfacing probability $\Pr(S = 1)$, utilitarian social welfare is

$$W(d) = \int_X \big(\mathbb{E}[S]\, u(x) - d(x)c(x)\big)\, \mathrm{d}F(x). \tag{23}$$

The utilitarian benchmark that ignores strategic participation is to choose S deterministically to maximize

$$\int_X S\, u(x)\, \mathrm{d}F(x).$$

[20] This is the standard pseudo-true-parameter interpretation of estimators under sample selection or misspecification (Heckman 1979).

Thus the welfare-maximizing rule is

$$S^* = \mathbb{1}\left\{\int_X u(x)\,\mathrm{d}F(x) \geq 0\right\}, \tag{24}$$

which surfaces the note if and only if its average utility in the whole population is non-negative.

By contrast, under Community Notes, the surfacing decision (22) is based on an estimate of *average approval among participants*, as in Proposition 7.2. If participation is selective—for example, because costs or perceived influence vary across types—the mechanism may implement a different rule than (24).

7.3.2 A one-dimensional ideology model

To make these forces concrete, consider a simple one-dimensional model of ideological heterogeneity. Let $x \in \mathbb{R}$ be a user's ideological ideal point, distributed with continuous density f that is symmetric about 0. Assume the note supports policy $+1$. User x's utility from surfacing the note is increasing in x:

$$u(x) = v(x), \qquad v'(x) > 0,$$

so "extreme" users (large $|x|$) care more intensely than moderates. We normalize $u(0) = 0$.

Each user who participates incurs a cost $c(x) \geq 0$, weakly increasing in $|x|$. This captures, for example, that extreme users may face higher opportunity costs or psychological costs of engaging with opposing content.

Community Notes aggregates ratings through a latent factor model, which we stylize as a weighted average of approval indicators $r(x) \in \{-1, +1\}$:

$$\hat{\theta}^{\mathrm{CN}} \approx \frac{\int_X w(x)\, d(x)\, r(x)\,\mathrm{d}F(x)}{\int_X w(x)\, d(x)\,\mathrm{d}F(x)},$$

where $w(x)$ are the mechanism's implicit weights on type x. For a bridging-based algorithm like Community Notes, the weights tilt toward users whose ratings are informative across diverse ideological neighborhoods (Wojcik, Hilgard, Judd, Mocanu, Ragain, Hunzaker, et al. 2022b), so it is natural to assume

$$w(x) = w(-x), \qquad w'(x) < 0 \text{ for } x > 0, \tag{25}$$

i.e. moderate users receive higher weight than extremes.

Given a belief about others' participation, each type x compares the expected benefit from rating to the cost. Let $\pi(x)$ denote the perceived probability that type x is pivotal for S (i.e. that their rating changes the sign of $\hat{\theta}^{\mathrm{CN}} - \tau$). Then the expected incremental payoff from rating for type x is approximately

$$B(x) \equiv \pi(x)\,\Delta u(x), \tag{26}$$

where $\Delta u(x)$ is the difference in utility between the outcomes $S = 1$ and $S = 0$, which is proportional

to $u(x)$. Costly voting models show that in large electorates, $\pi(x)$ is small and strongly shaped by the voting rule and weights (Palfrey and Rosenthal 1985; Börgers 2004). In the present setting, the bridging weights (25) imply that extreme types have both lower weights and lower pivot probabilities than moderates, so $\pi(x)$ is decreasing in $|x|$.

Assumption 7.3 (Monotone participation incentives). For each note, the net surplus from participating,

$$\Phi(x) \equiv B(x) - c(x),$$

is symmetric and strictly decreasing in $|x|$, with $\Phi(0) > 0$ and $\lim_{|x|\to\infty} \Phi(x) < 0$.

Assumption 7.3 is a stylized way of imposing that (i) all types strictly prefer the note if it is sufficiently aligned with their ideology (so $u(x)$ increases in x), (ii) bridging weights and pivot probabilities decrease in $|x|$, and (iii) costs do not fall fast enough with $|x|$ to offset the declining influence of extreme users. It is consistent with standard pivotal-voter models of costly voting (Palfrey and Rosenthal 1985; Börgers 2004).

Proposition 7.4 (Centrist participation and centrist filtering)**.** *Under Assumption 7.3, the costly-participation game for a fixed note has a symmetric threshold equilibrium characterized by a cutoff $t \geq 0$ such that:*

1. *The equilibrium participation set is*

$$d^*(x) = \mathbb{1}\{|x| \leq t\},$$

   so only moderate types rate the note.

2. *The Community Notes estimator converges to*

$$\theta_{\text{eq}}^{\text{CN}} = \mathbb{E}[Y \mid |X| \leq t],$$

   the average approval in the moderate subpopulation.

3. *The induced surfacing rule is*

$$S^{\text{CN}} = \mathbb{1}\left\{\int_{|x|\leq t} u(x)\,\mathrm{d}F(x) \geq 0\right\},$$

   which in general differs from the utilitarian rule (24). In particular, if extreme types have sufficiently large $|u(x)|$ for $|x| > t$, there exist notes for which $S^{\text{CN}} \neq S^*$.

Proof sketch. Given Assumption 7.3, best responses are monotone in $|x|$: if type x finds it optimal to participate, then any more moderate type x' with $|x'| < |x|$ has strictly higher net surplus $\Phi(x')$ and thus also participates. By symmetry of f and Φ, this yields a symmetric cutoff t. Standard

arguments for global games and costly voting then imply existence and uniqueness of such a threshold equilibrium.[21]

Part (ii) follows directly from Proposition 7.2: conditional on equilibrium participation $d^*(\cdot)$, Community Notes estimates the average approval among participants, which is exactly $\mathbb{E}[Y \mid |X| \le t]$. Part (iii) compares the equilibrium surfacing rule with the utilitarian rule. Because $u(x)$ grows in $|x|$ and extreme types are excluded by the cutoff, it is straightforward to construct examples where

$$\int_{|x|\le t} u(x)\,\mathrm{d}F(x) < 0 \quad \text{but} \quad \int_X u(x)\,\mathrm{d}F(x) > 0,$$

or vice versa, by concentrating sufficient utility mass in the tails of F. In such cases $S^{\mathrm{CN}} \neq S^*$ and the mechanism acts as a "centrist filter" that prioritizes moderate views over the intensity of extreme preferences. □

Proposition 7.4 shows that when participation is costly and extreme users perceive their influence as low, Community Notes can endogenously produce a more centrist outcome than a utilitarian benchmark. This is not necessarily undesirable—bridging-based weighting is explicitly designed to highlight notes that appeal across ideological lines—but it clarifies that the mechanism optimizes a notion of approval among participating moderates rather than total welfare.

7.4 Comparison to simple approval voting

Finally, we compare Community Notes to a simple approval-voting rule on notes. Under simple approval voting, each participant reports an approval indicator $A_{im} \in \{0, 1\}$ for note m, and the platform surfaces the note if the sample mean approval exceeds a threshold:

$$\hat{\theta}_m^{\mathrm{AV}} \equiv \frac{1}{S_m} \sum_{i \in R_m} A_{im}, \qquad S_m = \sum_{i=1}^{I} D_{im},$$

and

$$S_m^{\mathrm{AV}} = \mathbb{1}\{\hat{\theta}_m^{\mathrm{AV}} \ge \tau\}.$$

7.4.1 No selection bias

Under the no-selection model (10)–(12), and with A_{im} a monotone transformation of Y_{im}, both simple approval voting and Community Notes are unbiased estimators of the population mean approval θ_m. However, Community Notes is asymptotically more efficient because it exploits the factor structure to remove rater-specific heterogeneity.

To see this, suppose $A_{im} = \mathbb{1}\{Y_{im} \ge 0\}$ and the distribution of Y_{im} has continuous density at

[21] See, for example, the analysis of symmetric Bayesian equilibria in costly voting models (Palfrey and Rosenthal 1985; Börgers 2004).

zero. Then for large samples, central limit arguments imply

$$\sqrt{S_m}\big(\hat{\theta}_m^{\mathrm{AV}} - \theta_m^A\big) \xrightarrow{d} \mathcal{N}\big(0,\, \theta_m^A(1-\theta_m^A)\big),$$

where $\theta_m^A \equiv \mathbb{E}[A_{im}]$. The variance depends on the full distribution of Y_{im}. By contrast, Proposition 7.1 ensures that a factor-based Community Notes estimator of θ_m attains the smaller variance σ_ε^2/S_m. Using $\hat{\theta}_m^{\mathrm{CN}}$ in the threshold rule (22) therefore yields more statistically precise surfacing decisions than simple approval voting, in the same sense that generalized least squares dominates ordinary least squares when the error structure is correctly specified.

7.4.2 Selection bias and welfare

Once participation is endogenous, both mechanisms inherit the selection bias highlighted in Proposition 7.2. In each case, the estimand becomes the average approval among participants:

$$\lim \hat{\theta}_m^{\mathrm{AV}} = \mathbb{E}[A_{im} \mid D_{im} = 1], \qquad \lim \hat{\theta}_m^{\mathrm{CN}} = \mathbb{E}[Y_{im} \mid D_{im} = 1].$$

Statistical efficiency alone no longer guarantees higher welfare. In the costly-participation setting of Proposition 7.4, Community Notes—by design—puts relatively more weight on moderates, lowering extreme users' pivot probabilities and thereby raising their effective participation costs. Simple approval voting, which weights all participants equally, may sustain a broader set of equilibrium participants, including some extreme users with strong preferences.

From a welfare perspective, simple approval voting and Community Notes thus trade off two considerations:

- *Statistical efficiency.* Conditional on a given set of participants, Community Notes uses matrix factorization to construct a more efficient estimator of average approval, improving the accuracy of binary surfacing decisions.
- *Participation incentives.* Because Community Notes de-emphasizes ratings that are not "bridging," it may reduce the perceived influence of extreme users and thereby discourage their participation. If extremes have large-magnitude utilities, the resulting centrist filter can move the mechanism away from the utilitarian benchmark (24), potentially lowering total welfare despite higher statistical efficiency.

References

Allen, Joseph N. L., Cameron Martel, and David G. Rand (2022). "Birds of a Feather Don't Fact-Check Each Other: Partisanship and the Evaluation of News in Twitter's Birdwatch Crowdsourced Fact-Checking Program". In: *PsyArXiv*. URL: https://osf.io/preprints/psyarxiv/57e3q.

Amatriain, Xavier, Josep M. Pujol, and Nuria Oliver (2011). "Data Mining Methods for Recommender Systems". In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. Boston, MA: Springer, pp. 39–71. DOI: 10.1007/978-0-387-85820-3_2.

Anatolyev, Stanislav and Anna Mikusheva (2021). "Limit Theorems for Factor Models". In: *Econometric Theory* 37.5, pp. 1034–1074. DOI: 10.1017/S0266466620000468.

Bai, Jushan (2003a). "Inferential Theory for Factor Models of Large Dimensions". In: *Econometrica* 71.1, pp. 135–171. DOI: 10.1111/1468-0262.00392.

— (2003b). "Inferential Theory for Factor Models of Large Dimensions". In: *Econometrica* 71.1, pp. 135–171. DOI: 10.1111/1468-0262.00392.

Barberà, Salvador (2001). "An introduction to strategy-proof social choice functions". In: *Social Choice and Welfare* 18.4, pp. 619–653. DOI: 10.1007/s003550100151.

Barberà, Salvador, Bhaskar Dutta, and Arunava Sen (2001). "Strategy-proof social choice correspondences". In: *Journal of Economic Theory* 101.2, pp. 374–394. DOI: 10.1006/jeth.2000.2782.

Barberà, Salvador, Faruk Gul, and Ennio Stacchetti (1993). "Generalized Median Voter Schemes and Committees". In: *Journal of Economic Theory* 61.2, pp. 262–289. DOI: 10.1006/jeth.1993.1069.

Barberà, Salvador, Jordi Massó, and Alejandro Neme (1999). "Maximal domains of preferences preserving strategy-proofness for generalized median voter schemes". In: *Social Choice and Welfare* 16.2, pp. 321–336. DOI: 10.1007/s003550050146.

— (2005). "Voting by committees under constraints". In: *Journal of Economic Theory* 122.2, pp. 185–205. DOI: 10.1016/j.jet.2004.05.006.

Barberà, Salvador, Hugo Sonnenschein, and Lin Zhou (1991). "Voting by Committees". In: *Econometrica* 59.3, pp. 595–609. DOI: 10.2307/2938220.

Björkegren, Daniel (2024). "Manipulation-Robust Prediction". In: *Working Paper*. URL: https://dan.bjorkegren.com/manipulation.pdf.

Börgers, Tilman (2004). "Costly Voting". In: *American Economic Review* 94.1, pp. 57–66. DOI: 10.1257/000282804322970706.

Buterin, Vitalik (Aug. 2023). *What Do I Think About Community Notes?* URL: https://vitalik.eth.limo/general/2023/08/16/communitynotes.html.

Casella, Alessandra and Luis Sanchez (2018). "Storable Votes and Quadratic Voting: An Experiment on Four California Propositions". Working paper. URL: https://econ.columbia.edu/wp-content/uploads/sites/32/2017/10/draft.8.25.2018.pdf.

Chuai, Yuwei, Moritz Pilarski, Thomas Renault, David Restrepo-Amariles, Aurore Troussel-Clément, Gabriele Lenzini, and Nicolas Pröllochs (2024). "Community-based fact-checking reduces the spread of misleading posts on social media". In: *arXiv preprint*. URL: https://arxiv.org/abs/2409.08781.

Chuai, Yuwei, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini (2024). "Did the Roll-Out of Community Notes Reduce Engagement with Misinformation on X/Twitter?" In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW2, 428:1–428:52. DOI: 10.1145/3686967.

Clinton, Joshua, Simon Jackman, and Douglas Rivers (2004). "The Statistical Analysis of Roll Call Data". In: *American Political Science Review* 98.2, pp. 355–370. DOI: 10.1017/S0003055404001194.

Clinton, Joshua D., Simon Jackman, and Douglas Rivers (2004). "The Statistical Analysis of Roll Call Data". In: *American Political Science Review* 98.2, pp. 355–370. DOI: 10.1017/S0003055404001194.

Community Notes (2023). *Ranking Notes*. Accessed 29 November 2025. URL: https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes.

— (2024). *Note Ranking Algorithm*. https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes. Accessed 30 November 2025.

Community Notes Team (2023). *Note Ranking Algorithm*. Accessed 29 November 2025. URL: https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes.

Dai, Xiaowu (2024). "Incentive-Aware Recommender Systems in Two-Sided Markets". In: *ACM Transactions on Recommender Systems* 2.4, pp. 1–33. DOI: 10.1145/3674158.

Drolsbach, Chiara Patricia and Nicolas Pröllochs (2022). "Diffusion of Community Fact-Checked Misinformation on Twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media*. URL: https://arxiv.org/abs/2205.13673.

Drolsbach, Chiara Patricia, Kirill Solovev, and Nicolas Pröllochs (2024). "Community Notes Increase Trust in Fact-Checking on Social Media". In: *PNAS Nexus* 3.7, pgae217. DOI: 10.1093/pnasnexus/pgae217.

Georgescu, Laura, James Fox, Anna Gautier, and Michael Wooldridge (2024). "Fixed-Budget and Multiple-Issue Quadratic Voting". In: *arXiv preprint*. eprint: 2409.06614. URL: https://arxiv.org/abs/2409.06614.

Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters (2016). "Strategic Classification". In: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*. ITCS '16. New York, NY: Association for Computing Machinery, pp. 111–122. DOI: 10.1145/2840728.2840730.

Heckman, James J. (1979). "Sample Selection Bias as a Specification Error". In: *Econometrica* 47.1, pp. 153–161. DOI: 10.2307/1912352.

Jin, Kun, Xueru Zhang, Vincent Conitzer, Fei Fang, and Yang Liu (2022). "Incentive Mechanisms for Strategic Classification and Regression". In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC '22. New York, NY: Association for Computing Machinery, pp. 560–561. DOI: 10.1145/3490486.3538241.

Jones, Isaiah, Brent Hecht, and Nicholas Vincent (2022). "Misleading Tweets and Helpful Notes: Investigating Data Labor by Twitter Birdwatch Users". In: *Companion Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*. DOI: 10.1145/3500868.3559461.

Koren, Yehuda, Robert Bell, and Chris Volinsky (2009). "Matrix Factorization Techniques for Recommender Systems". In: *Computer* 42.8, pp. 30–37. DOI: 10.1109/MC.2009.263.

Lalley, Steven P. and E. Glen Weyl (2018). "Quadratic Voting: How Mechanism Design Can Radicalize Democracy". In: *AEA Papers and Proceedings* 108, pp. 33–37. DOI: 10.1257/pandp.20181002.

Lewis, Jeffrey B., Keith T. Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet (2025). *Voteview: Congressional Roll-Call Votes Database*. URL: https://voteview.com/.

Li, Haiwen, Soham De, Manon Revel, Andreas Haupt, Brad Miller, Keith Coleman, Jay Baxter, Martin Saveski, and Michiel A. Bakker (2025). "Scaling Human Judgment in Community Notes with LLMs". In: *Journal of Online Trust and Safety* 3.1, pp. 1–27. URL: https://tsjournal.org/index.php/jots/article/view/255.

Moulin, Hervé (1980). "On strategy-proofness and single peakedness". In: *Public Choice* 35, pp. 437–455. DOI: 10.1007/BF00128122.

Noema Magazine (2020). *Participation At Scale Can Repair the Public Square*. URL: https://www.noemamag.com/participation-at-scale-can-repair-the-public-square.

Ovadya, Aviv (2022a). *Bridging-Based Ranking*. Harvard Kennedy School Belfer Center for Science and International Affairs. URL: https://ash.harvard.edu/resources/bridging-based-ranking.

— (2022b). *Bridging-Based Ranking: How Platform Recommendation Systems Might Reduce Division and Strengthen Democracy*. Tech. rep. Belfer Center for Science and International Affairs, Harvard Kennedy School. URL: https://www.belfercenter.org/publication/bridging-based-ranking.

Ovadya, Aviv and Luke Thorburn (2023). "Bridging Systems: Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance". In: *arXiv preprint*. DOI: 10.48550/arXiv.2301.09976. eprint: 2301.09976. URL: https://arxiv.org/abs/2301.09976.

Palfrey, Thomas R. and Howard Rosenthal (1985). "Voter Participation and Strategic Uncertainty". In: *American Political Science Review* 79.1, pp. 62–78. DOI: 10.2307/1956119.

Pilarski, Moritz, Chiara Patricia Drolsbach, Gabriele Lenzini, and Nicolas Pröllochs (2023). "How Community Fact-Checkers Select Their Targets on Twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media*. URL: https://arxiv.org/abs/2305.09519.

Poole, Keith T. and Howard Rosenthal (1985a). "A Spatial Model for Legislative Roll Call Analysis". In: *American Journal of Political Science* 29.2, pp. 357–384. DOI: 10.2307/2111172.

— (1985b). "A Spatial Model for Legislative Roll Call Analysis". In: *American Journal of Political Science* 29.2, pp. 357–384. DOI: 10.2307/2111172.

— (1997a). *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press. URL: https://books.google.com/books?id=x9vyGly7hLcC.

— (1997b). *Congress: A Political-Economic History of Roll Call Voting*. New York, NY: Oxford University Press. URL: https://global.oup.com/academic/product/congress-9780195055771.

Pröllochs, Nicolas (2022). "Community-Based Fact-Checking on Twitter's Birdwatch Platform". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16, pp. 794–805. DOI: 10.1609/icwsm.v16i1.19335.

Saeed, Muhammad, Orestis Papakyriakopoulos, and Claudia Müller-Birn (2022). "Crowdsourced Fact-Checking at Twitter: How Does the Crowd Compare with Experts?" In: *Proceedings of the International AAAI Conference on Web and Social Media*. URL: https://arxiv.org/abs/2208.09214.

Salganik, Matthew J. and Karen E. C. Levy (2015). "Wiki Surveys: Open and Quantifiable Social Data Collection". In: *PLOS ONE* 10.5, e0123483. DOI: 10.1371/journal.pone.0123483.

Slaughter, Beth, Sam Rubin, Manoj Prasad, Kevin Kao, Aamir Memon, Raghav Ramaswami, Lada Adamic, and David Rothschild (2025). "Community Notes Reduce Engagement with and Diffusion of Misinformation on X". In: *Proceedings of the National Academy of Sciences* 122.X, e2503413122. DOI: 10.1073/pnas.2503413122.

Slaughter, Isaac, Axel Peytavin, Johan Ugander, and Martin Saveski (2025a). "Community Notes Reduce Engagement with and Diffusion of False Information Online". In: *Proceedings of the National Academy of Sciences* 122.38, e2503413122. DOI: 10.1073/pnas.2503413122.

— (2025b). "Community Notes Reduce Engagement with and Diffusion of False Information Online". In: *Proceedings of the National Academy of Sciences* 122.38, e2503413122. DOI: 10.1073/pnas.2503413122.

Small, Christopher, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill (2021). "Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces". In: *Recerca* 26.2, pp. 1–26. DOI: 10.6035/recerca.5516.

Stock, James H. and Mark W. Watson (2002). "Forecasting Using Principal Components from a Large Number of Predictors". In: *Journal of the American Statistical Association* 97.460, pp. 1167–1179. DOI: 10.1198/016214502388618960.

Sundaram, Ramesh, Anil Vullikanti, Haifeng Xu, and Yilin Yao (2023). "PAC-Learning for Strategic Classification". In: *Journal of Machine Learning Research* 24.223, pp. 1–52. URL: https://www.jmlr.org/papers/volume24/21-1250/21-1250.pdf.

The Computational Democracy Project (2023). *Polis: A platform for large-scale online deliberation*. URL: https://compdemocracy.org/polis/.

Twitter, Inc. (Jan. 25, 2021). *Introducing Birdwatch, a community-based approach to misinformation*. URL: https://blog.x.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

— (Oct. 6, 2022). *Helpful Birdwatch notes are now visible to everyone on Twitter in the US*. URL: https://blog.x.com/en_us/topics/product/2022/helpful-birdwatch-notes-now-visible-everyone-twitter-us.

UK Policy Lab (Oct. 11, 2022). *Cutting through complexity using collective intelligence*. URL: https://openpolicy.blog.gov.uk/2022/10/11/cutting-through-complexity-using-collective-intelligence/.

Warden, Jonathan (2024a). *Multidimensional Community Notes*. https://jonathanwarden.com/multidimensional-community-notes.

— (Jan. 2024b). *Understanding Community Notes and Bridging-Based Ranking*. URL: https://jonathanwarden.com/understanding-community-notes.

— (2024c). *Understanding Community Notes and Bridging-Based Ranking*. https://jonathanwarden.com/understanding-community-notes/.

— (2024d). *Understanding Community Notes and Bridging-Based Ranking*. URL: https://jonathanwarden.com/understanding-community-notes (visited on 11/29/2025).

Wikipedia contributors (2023). *Community Notes — X's community-based fact-checking system*. URL: https://en.wikipedia.org/wiki/Community_Notes.

Wirtschafter, Vivian et al. (2023). "Who Fact-Checks the Facts? Evidence from Twitter's Community Notes". In: *Journal of Online Trust and Safety* 2.1. URL: https://tsjournal.org/index.php/jots/article/view/139/57.

Wojcik, Stefan, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter (2022). "Birdwatch: Crowd Wisdom and Bridging Algorithms Can Inform Understanding and Reduce the Spread of Misinformation". In: *arXiv preprint*. DOI: 10.48550/arXiv.2210.15723. arXiv: 2210.15723. URL: https://arxiv.org/abs/2210.15723.

Wojcik, Stefan, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter (2022a). "Birdwatch: Crowd Wisdom and Bridging Algorithms Can Inform Understanding and Reduce the Spread of Misinformation". In: *arXiv preprint*. DOI: 10.48550/arXiv.2210.15723. eprint: 2210.15723. URL: https://arxiv.org/abs/2210.15723.

— (2022b). *Birdwatch: Crowd Wisdom and Bridging Algorithms Can Inform Understanding and Reduce the Spread of Misinformation*. arXiv: 2210.15723. URL: https://arxiv.org/abs/2210.15723.

X Corp. (2023a). *Community Notes*. URL: https://communitynotes.x.com.

— (2023b). *Community Notes open-source repository*. URL: https://github.com/twitter/communitynotes.

Yu, Yi, Tengyao Wang, and Richard J. Samworth (2015). "A Useful Variant of the Davis–Kahan Theorem for Statisticians". In: *Biometrika* 102.2, pp. 315–323. DOI: 10.1093/biomet/asv008.