From Birdwatch to Community Notes, from Twitter to X: four years of community-based content moderation

Saeedeh Mohammadi[1,2], Narges Chinichian[3,4,8], Hannah Doyal[4,8], Kristina Skutilova[5], Hao Cui[2], Michele d'Errico[2], Siobhan Grayson[6], and Taha Yasseri[1,2,7,9*]

[1]School of Mathematics and Statistics, University College Dublin, Dublin, Ireland
[2]School of Social Sciences and Philosophy, Trinity College Dublin, Dublin, Ireland
[3]Institute for Theoretical Physics, Technical University of Berlin, Berlin, Germany
[4]SPICED Academy, Berlin, Germany
[5]School of Computer Science, University College Dublin, Dublin, Ireland
[6]School of Sociology, University College Dublin, Dublin, Ireland
[7]Faculty of Arts and Humanities, Technological University Dublin, Dublin, Ireland
[8]These authors contributed equally
[9]Lead contact
[*]Correspondence: taha.yasseri@tcd.ie

SUMMARY

Community Notes (formerly known as Birdwatch) is the first large-scale crowdsourced content moderation initiative that was launched by X (formerly known as Twitter) in January 2021. As the Community Notes model gains momentum across other social media platforms, there is a growing need to assess its underlying dynamics and effectiveness. This Resource paper provides (a) a systematic review of the literature on Community Notes, and (b) a major curated dataset and accompanying source code to support future research on Community Notes. We parsed Notes and Ratings data from the first four years of the program and conducted language detection across all Notes. Focusing on English-language Notes, we extracted embedded URLs and identified discussion topics in each Note. Additionally, we constructed monthly interaction networks among the Contributors. Together with the literature review, these resources offer a robust foundation for advancing research on the Community Notes system.

KEYWORDS

Content Moderation, Community Notes, Birdwatch, Network Analysis, Topic Modelling

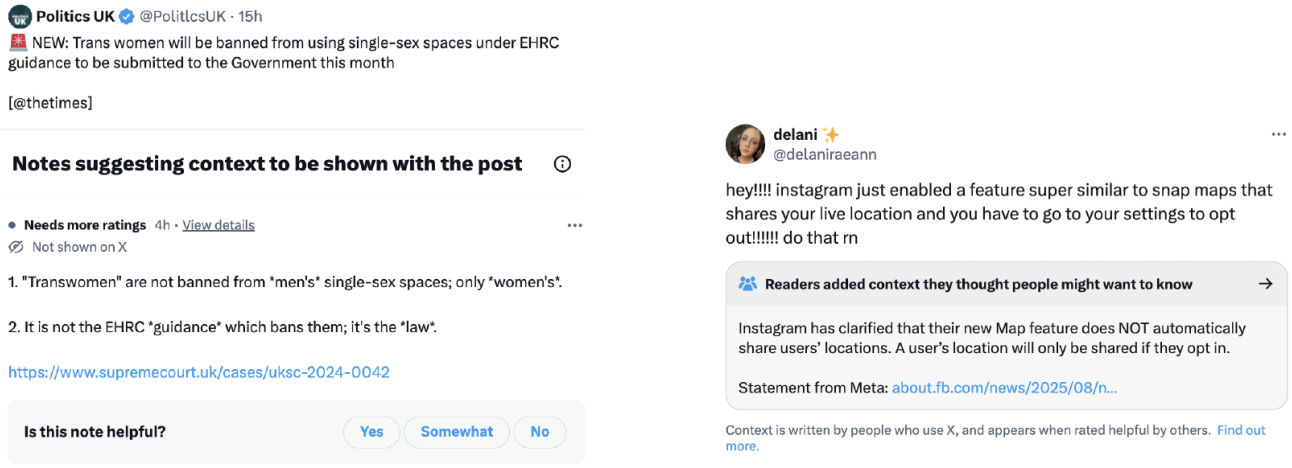
INTRODUCTION

The rapid production and spread of user-generated content on social media platforms in the absence of editorial oversight has increased users' exposure to false or misleading information[1]. In response, social media platforms have taken various approaches to content moderation. Content Moderation refers to the process of monitoring, flagging, or removing content that violates community guidelines or is deemed harmful. The main moderation strategies have been expert review and algorithmic systems. However, each comes with limitations. Expert evaluation, while often accurate, is costly and impractical at scale, given the sheer volume of content that needs moderation[2]. Automated algorithmic methods, on the other hand, are constrained by the biases and quality of their training data, often reinforcing systemic biases in classification[3]. In light of these challenges, community-based content moderation, leveraging the collective judgment of users, has emerged as a promising alternative.

On January 23, 2021, X (formerly Twitter) launched Community Notes (formerly Birdwatch), the first large-scale community-driven initiative to moderate misleading content. To participate, Contributors must have been active on X for at least six months, have a verified phone number, and maintain a clean record of rule compliance[4].

Contributors can write a Note to provide the missing context in relation to the Posts they find misleading. Other Contributors then rate these Notes as helpful, somewhat helpful, or unhelpful (Figure 1 (a)). The Note that receives the Helpful status based on the ratings from a diverse group of Contributors is displayed beneath the original Post in the timeline (an example is shown in Figure 1 (b))[5].

Each Contributor has two impact metrics: writing impact and rating impact. The former increases when a Contributor's Notes are consistently rated as helpful. The latter increases when the Contributor's rating of a Note aligns with the eventual status of the Note based on the rating algorithm, and decreases otherwise[6]. New Contributors begin by rating existing Notes and can only write their own Notes once their rating score reaches a threshold of 5[7]. Afterwards, the Contributors can decide to write a Note on any Post. Furthermore, users on X can request a Note for a Post; if a Post receives enough requests, top Contributors, those with high writing impact, are notified to draft a Note for that Post[8].



(a) (b)

Figure 1: **Examples of Notes.** (a) A Note that requires more ratings and is shown only to Community Notes Contributors for additional rating. (b) A Note that has received the status of helpful and is publicly shown beneath a Post.

Community Notes has experimented with various rating systems and decision rules to determine which note should be displayed for each post. The description provided above outlines only the core functionality of the system as of the time of writing; we do not aim to detail the specific algorithms and configurations, as these continue to evolve. In the pilot version of Community Notes, Notes that received the highest number of helpful ratings were picked to be shown in the timeline, regardless of the raters' past behaviour. This was not immune to polarised rating among Contributors[9] and partisan behaviour, such as Contributors rating Notes written by co-partisans as helpful and labelling Posts from cross-partisan Contributors as misleading[10]. To mitigate this issue, Community Notes adopted a *bridging algorithm*[11] in the rating system. This approach places Contributors along an opinion spectrum based on their historical rating behaviour. A Note is deemed helpful only if it receives enough positive ratings from Contributors with opposing viewpoints. The goal of this algorithm is to elevate Notes that are rated as helpful by a diverse set of Contributors. Community Notes is entirely open-source, with both the code and data available to the public[12].

Several studies have assessed the effectiveness of Community Notes in comparison with other content moderation methods. Findings indicate that it is effective at identifying misinformation, frequently flagging misleading content before expert fact-checkers. There is substantial agreement between Community Notes Contributors and professional fact-checkers, with their

classification aligning closely with expert evaluations[13]. Compared to expert-led moderation, Community Notes is also more scalable and cost-effective[14]. Unlike other crowdsourced approaches, such as "snoping", in which users respond directly to Posts with fact-checks, Community Notes Contributors tend to prioritise high-visibility content, particularly posts from influential users. When both systems evaluate the same content, a relatively rare occurrence, their assessments typically align[15].

Despite these advantages, Community Notes has notable limitations. It often operates more slowly than approaches such as "snoping"[15]. A recent study analysing over 2.2 million Posts found that 99.3% of misleading Posts received debunking comments within two hours of publication, whereas it takes an average of 24.29 hours for a Community Note to appear beneath a Post[16]. Moreover, Community Notes' open design leaves it vulnerable to manipulation and adversarial attacks[13]. It also relies heavily on the work of professional fact-checkers: one in twenty Notes explicitly references fact-checking sources, with this proportion increasing for sensitive topics[17]. These findings indicate that the production of high-quality Notes remains closely dependent on the broader fact-checking ecosystem.

While the studies above demonstrate that Community Notes can accurately identify misleading content and produce high-quality Notes, a critical question remains: can Community Notes reduce user engagement with such content? Results from A/B testing indicate that users exposed to Community Notes are 25–34% less likely to like or share flagged Posts[11]. Additional studies report that Posts with attached Notes have higher deletion rates[18], and that attached Notes reduce reposts. This effect is more pronounced for Posts containing embedded media than for text-only Posts[19]. In addition to reposts, an attached Note reduces the number of likes, views, and replies of a Post. Furthermore, when Notes appear beneath a Post, that Post tends to spread less widely and deeply across the platform[20].

Community Notes also shape how readers respond to misinformation. Kankham and Hou (2024) find that Notes are particularly effective in reducing belief and the spread of "wish" rumours (misinformation aligned with users' desires), whereas presenting related news articles is more effective in countering "dread" rumours (misinformation that evokes fear)[21]. Users generally view Community Notes more favourably than other interventions, such as misinformation flags. Unlike binary labels that simply mark content as false, Notes provide explanatory context,

which fosters greater trust and acceptance. Drolsbach et al. (2024) show that this contextual approach is more persuasive, as users tend to prefer interpretive frameworks over authoritative declarations of truth[22].

However, recent studies indicate that the overall rollout of Community Notes has not substantially curtailed engagement with misleading content. This limitation is attributed mainly to the delay in Notes reaching "helpful" status, often well after a Post's most viral phase[23,24]. Renault et al. (2024) report that while annotated Posts experience nearly a 50% drop in reposts and over a 30% reduction in replies and quotes, this effect is highly temporal. On average, it takes 15 hours for a Note to be published, by which time a Post has typically reached 80% of its total audience[18]. Similarly, Chuai et al. (2024) report an even longer delay—an average of 75.5 hours between post creation and note attachment—by which time 96.7% of reposts have already occurred[19]. De et al. (2024) report that in 91% of Posts where at least one Note was proposed, none reached "helpful" status[25]. This delay is particularly problematic in time-sensitive or politically charged contexts. For instance, one report found that 74% of accurate Notes related to the 2024 U.S. presidential election were never shown to users, allowing misleading Posts without Notes to spread 13 times faster than those annotated with Community Notes[26]. Researchers attribute this to the design of the rating algorithm: the more polarising the content (e.g., national elections), the less likely accurate Notes are to receive "helpful" status[27]. Related to this is the report that suggests most contributors have yet to produce a single Note rated as helpful[28].

Community Notes influence both posting behaviour and the ways users engage with content on the platform. Studies find that users whose content receives Notes tend to Post less overall, yet their subsequent Posts exhibit greater cognitive processing. Additionally, exposure to Community Notes improves reliability in users' own writing, while reducing extreme sentiment. However, these cognitive benefits are accompanied by decreased overall activity, as users exposed to Community Notes participate less frequently[29]. Receiving a Note can also have the unintended effect of increasing a user's visibility: users, particularly those with smaller followings, who are "noted" often gain additional followers, potentially amplifying their reach rather than diminishing it[28].

Several studies have examined the characteristics of Notes produced in Community Notes to understand better what makes a Note more effective or persuasive. Research indicates that

Notes on Posts that are flagged as misleading tend to be longer, more complex, and more negatively worded; traits that may make them more informative but potentially less accessible to general audiences[30]. The emotional tone of a Note also plays a key role: while emotionally charged language can increase engagement, it tends to reduce perceived trustworthiness. On average, Notes with high emotional content are rated as less helpful than those adopting a more neutral tone[31]. Moreover, Notes that closely align with the topic of the associated Post receive higher helpfulness ratings[32].

Citing sources is another critical element of Note quality. Contributors are encouraged to support their claims with external references, and including a source increases the likelihood of a Note reaching "helpful" status by a factor of 2.33. However, sources perceived as politically biased reduce the likelihood of a Note being rated as helpful[33]. Kangur et al. (2024) analysed citation patterns and found that the most frequently referenced sources were X and Wikipedia, with a noticeable left-leaning bias. Notes that cited more balanced or factually rigorous sources received higher helpfulness scores, reaffirming the other findings[34].

Contributors select which Posts to write Community Notes for based on multiple factors. A recent study shows that among 90,000 Posts for which X users requested a Note, the Posts perceived as more misleading by GPT-4.1 and the Posts by authors who are more frequently fact-checked, were more likely to attract Notes[35]. Political partisanship also plays a central role in how Contributors both select and evaluate Notes. Prior research shows that partisanship is often a stronger predictor of judgment than the content itself. Contributors are more likely to evaluate Posts from political opponents negatively and to rate Notes written by Contributors of opposing affiliations as unhelpful, thereby reinforcing existing ideological divides[10]. Network analyses of positive interactions reveal that the initial rating system, where the Note with the highest number of helpful ratings was surfaced, did not promote cross-partisan support. Instead, Contributors clustered into polarised groups[9]. This pattern mirrors broader trends of political polarisation observed on social media. Studies of signed interaction networks in Community Notes reinforce this finding, showing that Contributors consistently form clusters based on shared political ideologies[36]. Topic modelling and network analyses further demonstrate that such polarisation is especially pronounced in political discussions, whereas evaluations of non-political Posts display substantially less ideological bias[37].

Despite this, there is evidence that diversity within the pool of fact-checkers can improve fact-checking outcomes. Research comparing individual tagging (e.g., labelling misinformation independently) with Community Notes finds that the latter exposes users to more diverse perspectives. Unlike individual tagging, which can deepen echo chambers, Community Notes' peer-review structure temporarily increases informational diversity[38]. Although this effect is not lasting, it underscores the potential of collective moderation to mitigate polarisation when thoughtfully implemented. Research suggests that Community Notes can improve the quality of public discourse and support misinformation detection, drawing parallels to deliberative platforms such as Polis, which structure participation to promote reflection and consensus[39].

The system's evolution has also influenced Contributors' behaviour over time. Early in the program, most Notes were written on Posts from Contributors estimated to lean liberal[28]. However, as the Contributor base broadened, the partisan distribution shifted. A 2025 study by Renault et al. found that Posts from Republican users were more likely to be flagged with Community Notes, reflecting ongoing changes in participation and focus[40].

While Community Notes offer scalable fact-checking and collective decision-making benefits, it also carries risks, including system manipulation, exploitation of unpaid labour, and the marginalisation of underrepresented voices[41]. Furthermore, some features—such as prompting Contributors to revisit flagged content—may unintentionally reinforce exposure to misinformation[42]. Other ethical concerns have also been raised about the open structure of Community Notes, particularly the potential psychological harm to Contributors who encounter harmful or sensitive content[43].

Several studies have proposed modifications to Community Notes to enhance its effectiveness. Lloyd et al. (2025)[44] argue that researchers should focus on how the design and implementation of such systems influence their efficiency, including features such as interaction between Contributors, Note presentation, and the integration of technological advances such as Large Language Models (LLMs). Various approaches have already been explored. HawkEye, for instance, is a graph-based algorithm that jointly evaluates the quality of Contributors, Notes, and Posts through repeated iterations. It applies basic scoring principles and a smoothing technique to address cases with limited data, thereby improving the reliability and consistency of its evaluations[45]. Supernotes adopt a different approach by leveraging large language models (LLMs)

to generate Notes that are more likely to be rated as “helpful” than those written by individual Contributors. In this framework, an LLM reads all existing Notes, generates multiple candidate Notes, and evaluates them with a Personalised Helpful Model (PHM), which estimates the probability of receiving a “helpful” rating from a synthetic jury of diverse raters. The highest-scoring Note is then presented as the Supernote, and empirical evidence shows that Contributors rate these as significantly more helpful than the best human-written Notes[25]. Some scholars extend this idea into a human–AI hybrid framework in which both human Contributors and LLMs can propose Notes, but only humans serve as raters and evaluators. This division of labour preserves the legitimacy of human judgment while enabling LLMs to accelerate the generation of high-quality Notes[46]. Pushing this idea further, Costabile et al. (2025) simulated crowds using generative agents with diverse demographics and ideological perspectives. These agents outperformed human crowds in truthfulness classification, demonstrated higher internal consistency, were less susceptible to social and cognitive biases, and relied more on systematically informative criteria[47].

A second line of research emphasises diversity and collaboration in the Note-writing process rather than focusing solely on the rating stage[9]. In an online experiment, Juncosa et al. found that duos produced more helpful Notes than individuals, with diverse duos performing particularly well on Republican-leaning Posts[48]. Building on this insight, Mohammadi and Yasseri (2025) used LLMs to synthesise political diversity during Note writing. In a study involving over 800 participants, Contributors who received argumentative feedback from an LLM were more likely to produce higher-quality Notes[49].

In this work, we collected all Community Notes and corresponding ratings published over a four-year period, from January 23, 2021, to January 23, 2025. We applied language detection to filter for English-language Notes and, on this subset, conducted topic modelling, URL and domain extraction, and constructed interaction networks between Contributors based on their ratings of each other’s Notes. All processed data—including annotated Notes, interaction networks, and analysis outputs—are publicly released to support future research (see *Resource Availability* section). Although our present analysis is limited to English, the accompanying codebase is readily adaptable for multilingual extensions.

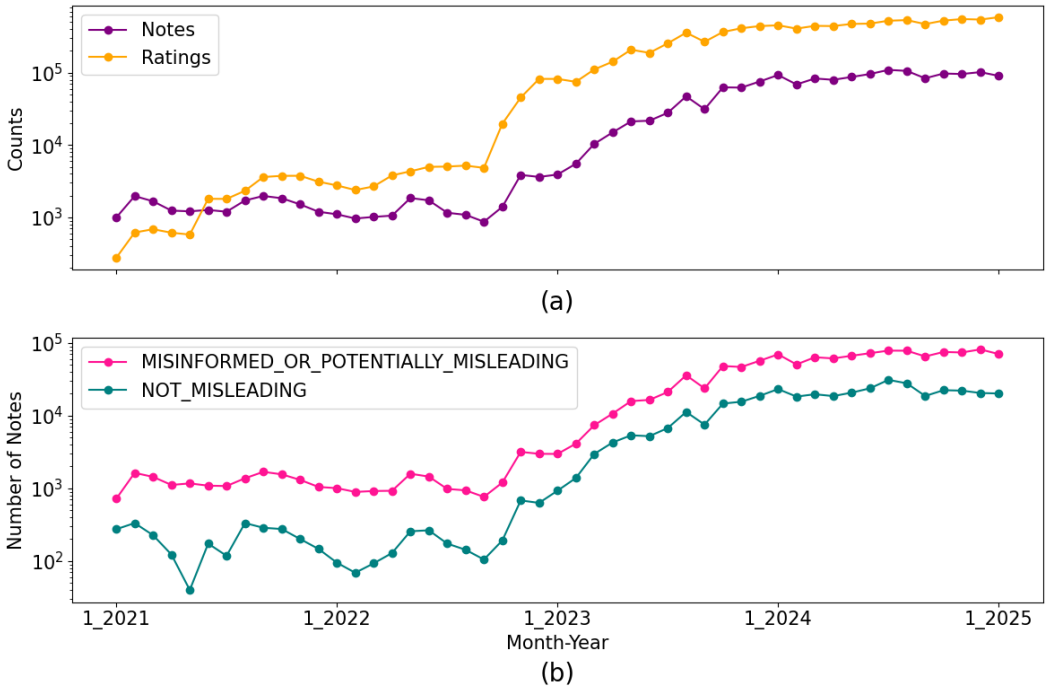


Figure 2: **Classification and the number of Notes per month.** (a) The number of Notes and Ratings created within each month on a log scale. (b) The monthly number of Notes that classify the Posts as ”potentially misleading or misinformed” and ”not misleading” over four years on a log scale.

RESULTS

The current design of Community Notes allows Contributors to classify a Post as either misleading or not misleading. The accompanying Notes serve to justify these classifications and provide additional context for the Post. Figure 2 (b) presents the number of Notes tagged as misleading or not misleading over the four-year period, revealing a consistent tendency for Contributors to label Posts as misleading more frequently.

Since its launch in January 2021, the Community Notes system has accumulated a substantial volume of contributions: 227,702 unique Contributors have written 1,614,743 Notes. However, a small number of contributors are responsible for a large share of the Notes. One Contributor alone has authored 33,186 Notes, an account that appears to be automated, consistently flagging impersonation attempts related to Non-Fungible Tokens (NFTs) and cryptocurrency scam accounts. The distribution of Notes authored per Contributor is highly skewed, as illustrated in Figure 3(a).

A similar pattern emerges in the distribution of Ratings. Figure 3(b) shows the rank-plot of the number of Ratings submitted by each Contributor. Such patterns are ubiquitous in online

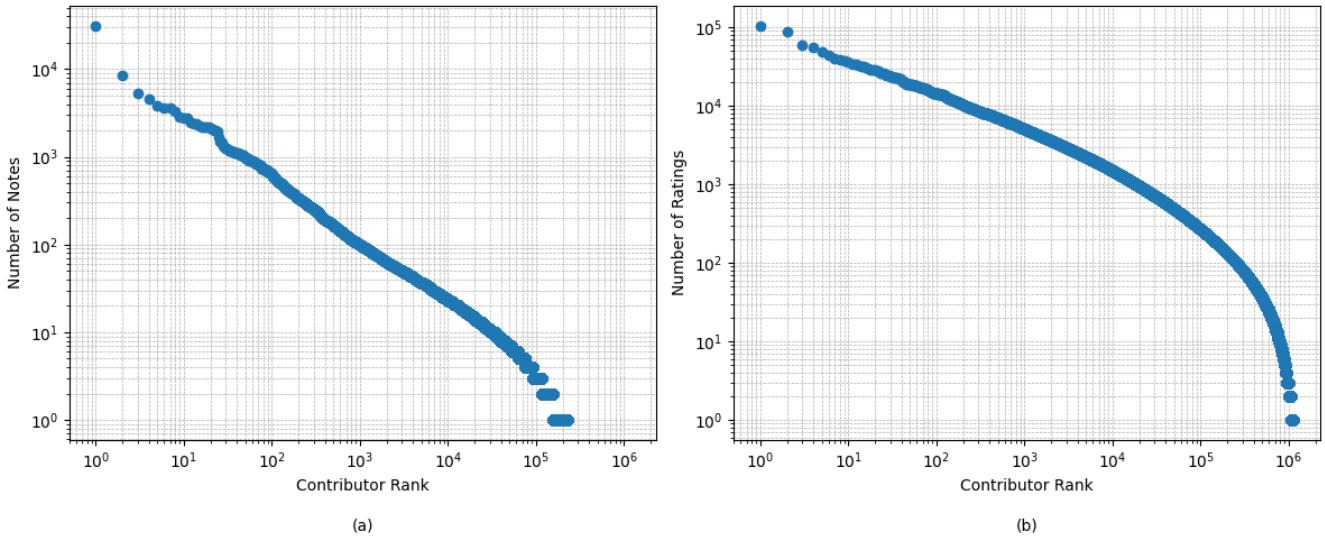


Figure 3: **The distribution of Notes and Ratings of Contributors.**(a) Log-log rank-plot showing the number of Notes written by each Contributor vs their rank. (b) Log-log rank-plot of the number of Ratings committed by each Contributor vs their rank.

open crowdsourcing platforms such as Wikipedia[50,51], citizen science projects[52], Reddit[53], and Urban Dictionary[54], where a small number of ”power users” contribute a large proportion of the activities. This concentration of users can challenge the practical realisation of representation and openness in these systems.

The Notes in our dataset were written on 1,016,673 distinct Posts. As shown in Figure 4(a), the distribution of Notes per Post is highly skewed: most Posts receive only a few Notes, while a small number attract substantial attention. The most annotated Post in the dataset has 90 Notes. This Post, made by Donald Trump on August 25, 2023, includes his mugshot alongside the caption: *”Election Interference, Never Surrender! DonaldjTrump.com”*. As of this writing, despite the large number of Notes on this Post, none of them have reached ”helpful” status, and therefore, no Note appears beneath it on the platform.

This pattern suggests a limitation in the current rating system. Highly visible Posts tend to attract many competing Notes. Although the number of ratings also increases with the number of Notes (Figure 4(d)), many of these Posts still do not have a Note displayed. A likely explanation that can be tested using the data provided in this Resource is as follows: when numerous Notes compete, Contributors select which Note to rate in line with their ideological preferences. As a result, no single Note gathers enough cross-perspective ratings to be marked as helpful, even though some Notes receive a large volume of ratings (Figure 4(b)).

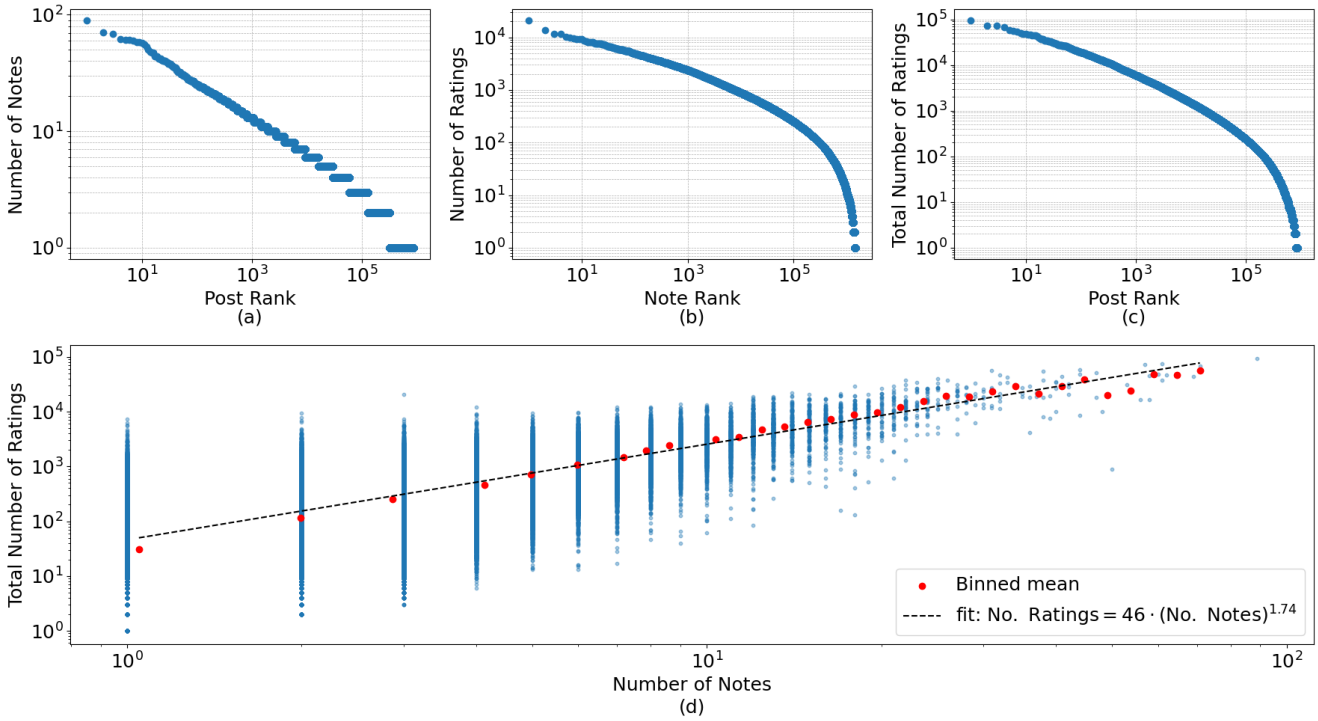


Figure 4: **The distribution of Notes and Ratings of Posts.** (a) Log-log rank plot of the number of Notes written on each Post. (b) Log-log rank plot of the number of Ratings each Note has received. (c) Log-log rank plot of the total number of Ratings on all Notes written on each Post. (d) Scatter plot (log-log) of the total number of Ratings on all Notes written on each Post versus the number of Notes written on that Post. The red dots represent the binned mean number of Ratings for a given range in the number of Notes. The black dashed line shows the fitted power-law relationship $y = A \cdot x^b$, where $A = 46$ is the scaling coefficient and $b = 1.74$ is the exponent.

As of December 7, 2023, Community Notes has been made available to X users in over 60 countries, with Notes written in 103 different languages. Figure 5 shows the top 10 languages in the dataset, along with the number of Notes written in each. The vast majority of Contributors write in only one language: only 35,515 Contributors, approximately 16% of all Note authors, have written Notes in more than one language. Even among the most multilingual Contributors, Notes are typically written in a single language. For example, the Contributor who used the largest number of distinct languages has written Notes in 17 different languages. Of their 500 Notes, 412 are in English, while most of their Notes in other languages consist of just a single Note each.

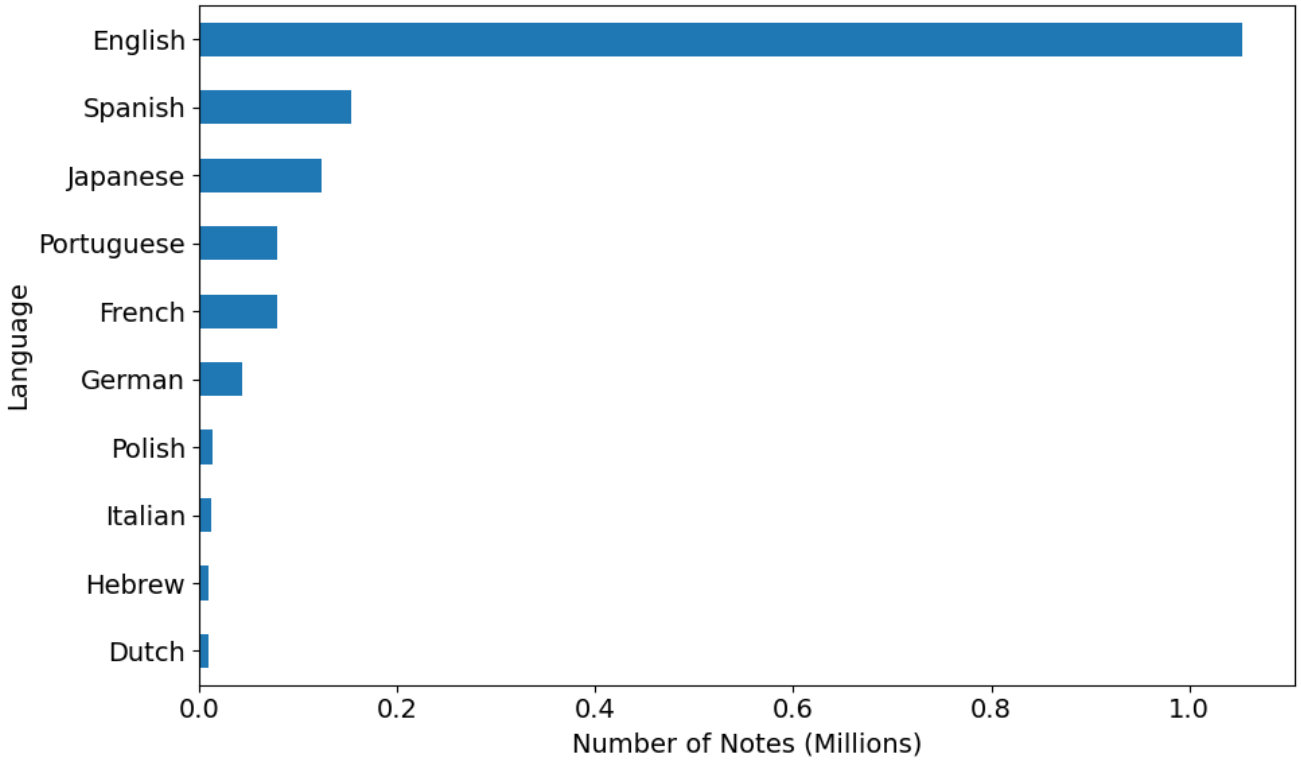


Figure 5: **Top 10 languages in Community Notes.** The number of Notes written in each of the top 10 languages used in Community Notes.

Given the prominence of English among Community Notes, the rest of our analysis focuses on this language. Among English Notes, 79.6% contain at least one URL, reflecting the platform's emphasis on source-backed claims. While drafting a Note, Contributors are instructed to include links to external sources and are prompted to indicate whether they believe the source would be considered trustworthy by most people.

We classified the 30 most frequently cited domains by political leaning using Ad Fontes Me-

dia's 2023 ratings. Ad Fontes assigns each domain a score on a left–neutral–right spectrum (Left: –20 to –10, Neutral: –9 to +9, Right: +10 to +20)[55]. Figure 6 displays the top 30 cited domains and the number of times each was referenced. Crowd-sourced platforms such as Wikipedia, X, and YouTube rank among the most frequently cited. The bar colours represent the (U.S.-based) political leaning of each domain, revealing that most cited sources are classified as neutral, although several left-leaning domains also appear.

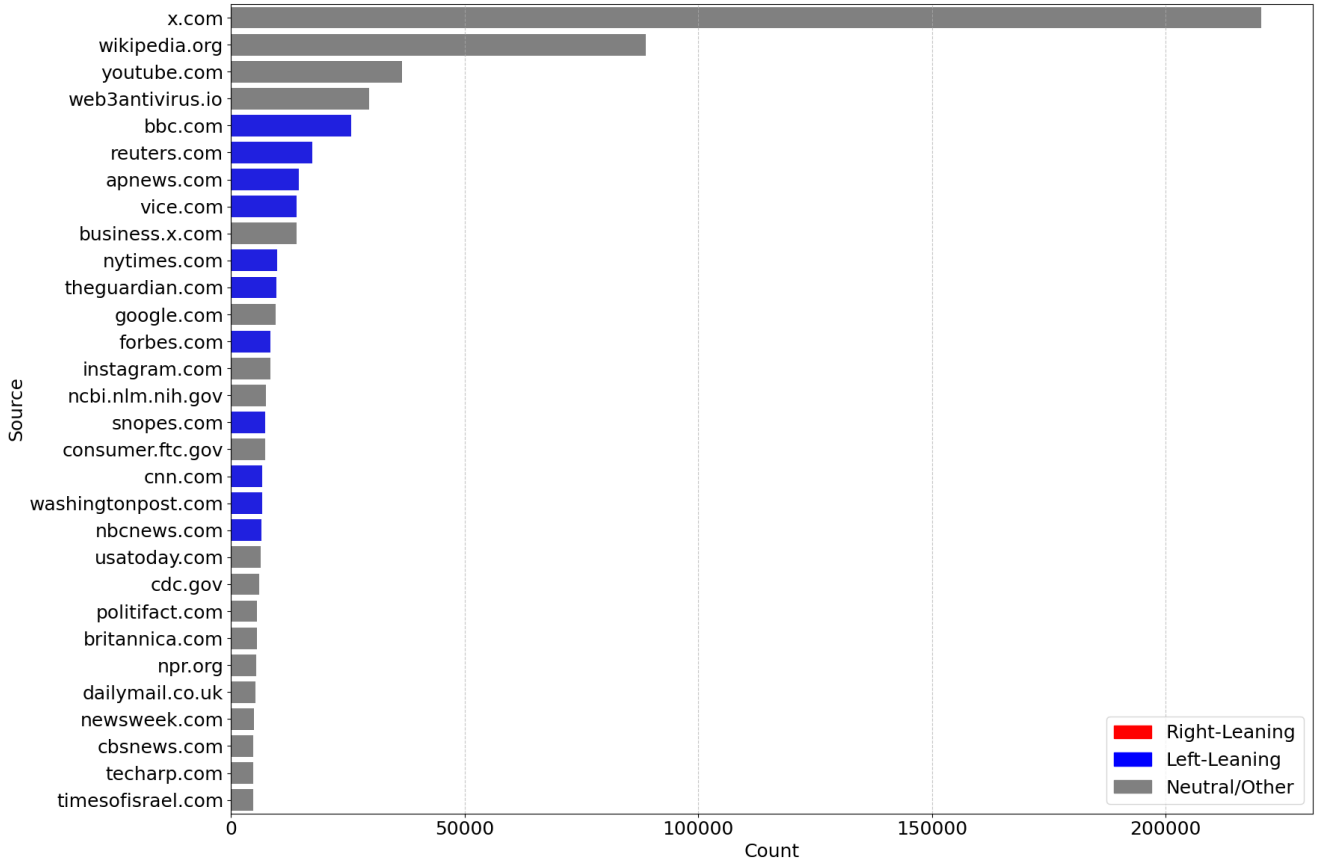


Figure 6: **Top 30 sources cited in Community Notes.** Top 30 cited domains in Community Notes, coloured by their (U.S.-based) political leaning. The X-axis shows the number of times each domain has been referenced in the Notes.

Among English-language Notes, we identified 57 distinct topics. Table 1 presents the ten largest topic clusters, excluding a noise cluster (see Methods). The most prominent themes include COVID-19, electoral processes, and ongoing geopolitical conflicts such as the Israel–Palestine and Russia–Ukraine wars, reflecting the strong connection between global events and Community Notes activity. Other topics, including cybersecurity, satire, misinformation, and social commentary, also emerge. This diversity underscores the broad scope of online discourse and the varied challenges it presents for content moderation. A complete list of topics and their repre-

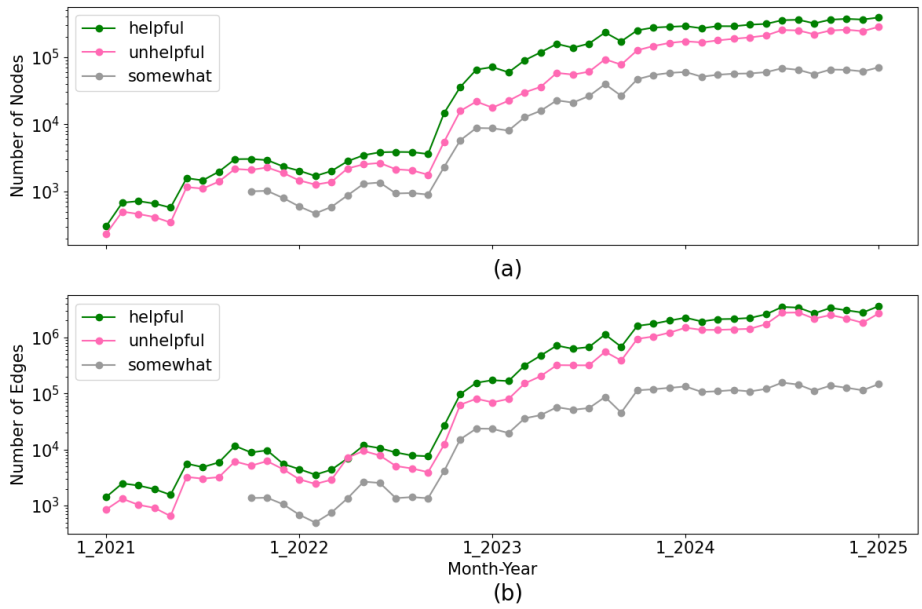


Figure 7: **The size of the Ratings network over time.** (a) The number of nodes for the helpful, unhelpful, and somewhat helpful networks for each month on the log scale. (b) The number of edges for the helpful, unhelpful, and somewhat helpful networks for each month on the log scale. Since the ”somewhat helpful” option was only introduced at the end of June 2021, the somewhat helpful network did not exist before that point.

sentative keywords is provided in Table S8, and the topic assignment to each Note is available in the accompanying dataset.

We constructed monthly rating networks covering the full four-year dataset. In each network, nodes represent Contributors, and directed edges represent Ratings: an edge from Contributor A to Contributor B indicates that A rated a Note written by B. The edge weight corresponds to the number of such ratings. For each month, we generated three separate networks: one for helpful ratings, one for somewhat helpful ratings, and one for unhelpful ratings. Figure 7 shows the number of nodes and edges in each of these networks over time.

Analysis of these rating networks offers valuable insight into the underlying community structure of the platform and its evolution over time (see, e.g., the work by Yasseri and Menczer (2023)[9]). As an illustrative example, Figure 8 presents visualisations of the three interaction networks—helpful, somewhat helpful, and unhelpful—constructed from ratings made in January 2023 as a representative month. This figure is intended primarily for illustration, and further analysis is required to fully interpret the structure of the networks. In particular, the unhelpful network (Figure 8(c)) warrants additional attention, since its edges represent negative interactions, yet they were treated as unsigned edges in the current visualisation.

Table 1: **Top 10 topics.**The top 10 topics with the highest number of Notes in English Community Notes and their representative keywords.

Topic Number	Number of Notes	Topic Name	Representative Words of Topic
56	192773	COVID-19 & Elections	vaccine, covid, video, image, vote, photo, trump, election, year, claim
44	79993	Community Feedback & Moderation	nnn, note, opinion, community, comment, abuse, post, personal, stop, need
28	76438	Israel-Palestine Conflict	israel, hamas, gaza, israeli, palestinian, jewish, attack, jews, palestine, terrorist
46	50638	Gender, Crime & Law Enforcement	woman, gender, sex, gun, crime, police, male, court, female, charge
34	49949	Cybersecurity & Crypto Scams	account, phishe, metamask, asset, ethereum, antivirus, password, malicious, detector, advisory
33	33628	Satire & Parody	joke, satire, satirical, clearly, parody, nnn, account, note, post, need
35	21820	Twitter Discussions	tweet, twitter, note, opinion, need, community, nnn, reply, original, express
37	20830	Russia-Ukraine War	ukraine, russia, russian, ukrainian, nato, putin, war, invasion, missile, invade
54	11011	Misinformation & Fact-Checking	claim, news, source, evidence, fact, false, article, debunk, fake, information
5	6908	Gambling & Advertising Violations	gambling, stake, undisclosed, contain, service, advertisement, promote, term, violate, tos

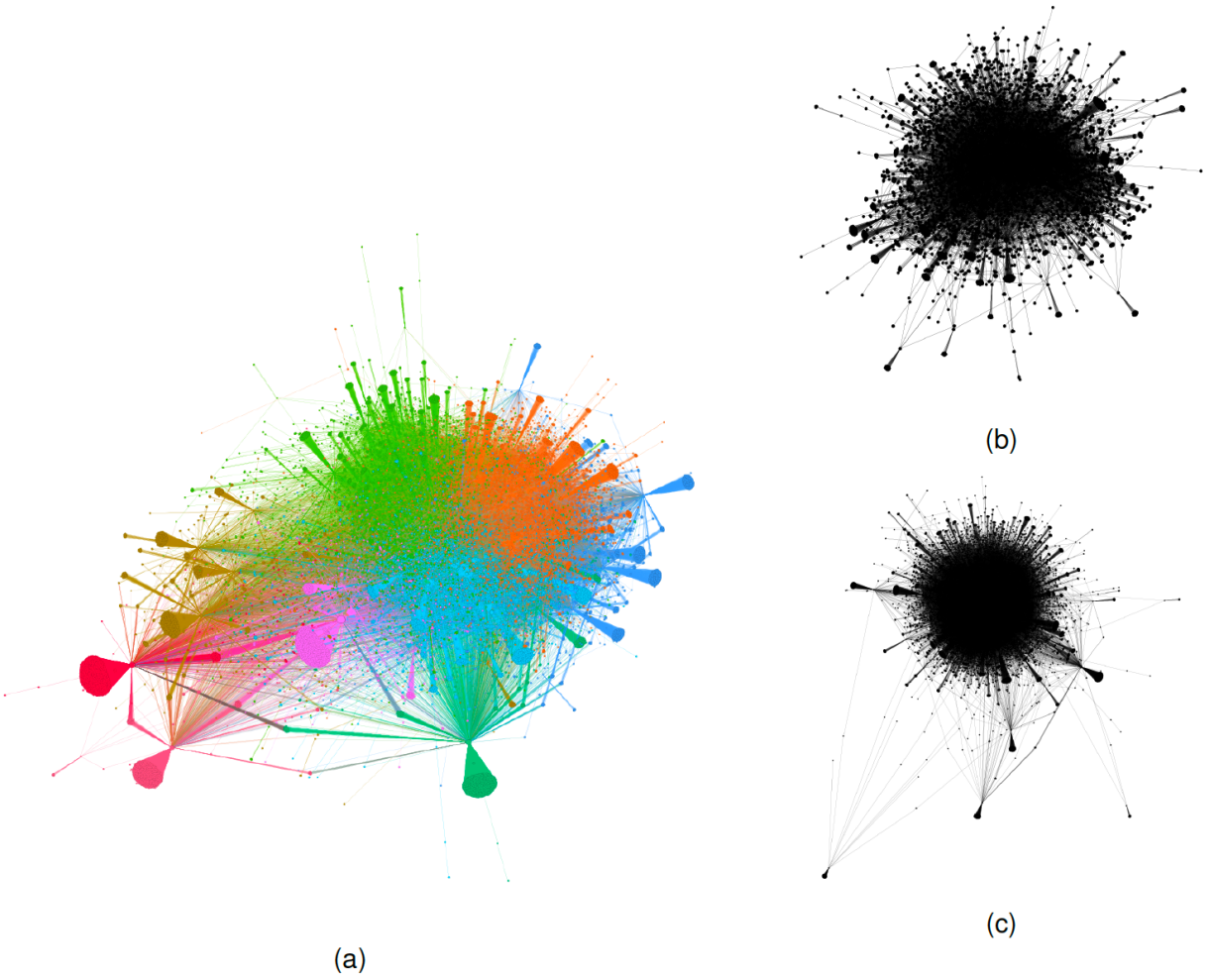


Figure 8: **Networks of January 2023.** The network visualisation of (a) helpful ratings, (b) somewhat helpful ratings, and (c) unhelpful ratings made in January 2023. Node colours indicate cluster membership, with clusters identified using the Louvain algorithm[56] with a Louvain resolution of 1. Only clusters containing more than 4% of all nodes are shown. In all networks, the size of each node is proportional to its weighted in-degree, representing the number of ratings a Contributor received during that month. The visualisations were created in Gephi 0.10.1.

Data Download

The Note IDs, along with the detected language of each Note, extracted URLs, associated domains, and detected topics, are available for download, covering the entire four-year period examined in this article.

For the ratings data, 49 TSV files are provided, each corresponding to a specific month as indicated by the filename. Accompanying each TSV file are four GraphML files. The first is a network file in which each directed edge between Contributors includes three attributes—helpful, unhelpful, and somewhat helpful—representing the number of times the source node rated Notes written by the target node with each rating type during that month. The remaining three GraphML

files are subgraphs of this full network, each isolating one rating type. In these subgraphs, edge weights indicate the number of times the source node rated the target node's Notes with the corresponding rating. All nodes are labelled with participant IDs across all network files. The code used to generate all datasets described above is also available for other researchers to use and extend.

DISCUSSION

The spread of misleading content on social media has had serious consequences, affecting public health outcomes[57,58], undermining electoral integrity[59,60], and even inciting violence[61]. In response, platforms have implemented a range of content moderation strategies, including third-party fact-checking and automated detection techniques. Introduced on X in January 2021, Community Notes is the first large-scale, community-driven content moderation system and has since begun gaining traction on other platforms as well[62,63].

Community Notes represents a fundamentally new approach to content moderation, distinguished not only by its crowdsourced nature but also by its epistemological foundation. As Augenstein et al. (2025) argue, there is a core philosophical distinction between Community Notes and traditional fact-checking systems[43]. Whereas traditional fact-checking is framed as the pursuit of an objective truth established through evidence, Community Notes are treated as subjective constructs reflecting personal opinion. This distinction has sparked debate over the functionality, credibility, and ethical implications of Community Notes as a moderation tool.

Community Notes leverages a broad pool of Contributors to evaluate a wide range of content and has shown promise in identifying misleading information[14] and providing accurate, context-rich Notes[13]. Compared to traditional fact-checking interventions, Community Notes are less intrusive, giving users the autonomy to engage with or disregard them[43]. Research also indicates that these Notes are perceived as more trustworthy than simple warning labels that lack explanation[22]. Furthermore, the system's algorithmic transparency has been shown to enhance user trust[43].

The democratic, consensus-based design of Community Notes is particularly valuable for addressing ambiguous claims, where a simple binary label—misleading or not—may oversimplify

the issue. In such cases, imposing a single authoritative judgment can backfire, fuelling distrust or even conspiracy thinking. By contrast, a crowd-sourced explanation that incorporates diverse perspectives can offer a more nuanced and credible interpretation, encouraging users to critically evaluate content rather than passively accept a top-down verdict.

Fact-checking complex or high-stakes Posts often demands domain-specific expertise that most Contributors lack, and even knowledgeable Contributors may never encounter such Posts[43]. This limitation is compounded by the platform's rating structure, which leaves a large share of Notes unpublished. Between January 23, 2021, and January 23, 2025, only 13.55% of Posts with at least one proposed Note ever received a "helpful" Note. Across the same period, 87.7% of all Notes remained in the "Needs More Ratings" category, and only 8.3% ultimately achieved "helpful" status and appeared beneath their respective Posts. Even then, the average delay was 26 hours—well past the point of peak visibility for most misleading Posts—greatly reducing the intervention's practical impact[64].

Several factors contribute to this bottleneck. First, ideological echo chambers limit exposure to diverse viewpoints, slowing the consensus-building needed for publication[43]. Second, political polarisation hinders agreement on contentious topics, making it harder for Notes to gain the cross-ideological support required to be deemed helpful[13].

Another key vulnerability of Community Notes is its susceptibility to manipulation by automated accounts. In our dataset, the Contributor with the highest number of authored Notes appears to be a bot-like account that consistently targets cryptocurrency-related Posts and exclusively links to a single website: `web3antivirus.io`. Consequently, this domain ranks as the fourth most cited source, surpassed only by X, Wikipedia, and YouTube. This observation underscores the presence of automated activity within the Community Notes ecosystem and raises concerns about the platform's exposure to coordinated manipulation.

Future research should further examine the limitations and vulnerabilities of Community Notes identified in this Resource article. To support such work, in addition to the systematic literature review presented above, we release a four-year dataset (January 23, 2021–January 23, 2025) containing every Note ID with its detected language, along with extracted topics, cited links, and domains for English-language Notes. We also provide monthly interaction networks capturing three rating types—helpful, somewhat helpful, and unhelpful—and the code needed

to reproduce or adapt the dataset. This Resource enables investigations into the functioning of Community Notes and the broader topic of online content moderation.

METHODS

For Note content, Community Notes provides a single TSV file containing all Notes, with metadata including the Note ID, author ID, referenced Post ID, creation timestamp, the Contributor's classification of the Post as misleading or not misleading, whether trustworthy sources were cited, and the full Note text[65].

Ratings are available in multiple TSV files, each entry representing a single rating. These records contain the Note ID, the rater's participant ID, perceived helpfulness, and reasons for the evaluation. The helpfulnessLevel field, introduced on June 30, 2021, replaced the earlier binary helpful/notHelpful label[65].

Although additional datasets—such as Note status history and Contributor enrolment records—are available, our analysis focuses exclusively on the Notes and Ratings datasets.

Content

We used the pre-trained FastText language identification model[66] to detect the language of each Note. English-language Notes were then processed through a custom text-cleaning pipeline optimised for topic modelling. This pipeline included HTML decoding, normalisation of accented characters, expansion of contractions, and removal of URLs, user mentions, hashtags, and non-alphabetic characters. We then removed stopwords, filtered out short tokens, and applied lemmatisation. Notes reduced to empty text after preprocessing were excluded to ensure clean input for embedding generation and topic modelling.

An important feature of Community Notes is the inclusion of URLs that provide supporting evidence for a Note's content. To detect these URLs, we implemented a script that extracted all links from each Note using regular expression (regex) patterns, as listed in Table S1. Extracted URLs were cleaned by removing whitespace and surrounding brackets. For domain-level analysis, we extracted root domains using a secondary regex (Table S2), which isolated the core domain and returned a list of extracted domains for each Note. We then normalised the extracted

domains to account for platform rebranding, URL shorteners, and regional variants. This step consolidated cases such as the 2023 rebranding of Twitter to X, the use of URL shorteners, and multi-region domain aliases (see Table S3).

We applied BERTopic[67] for topic modelling. BERTopic is a modular framework comprising sequential stages, each customisable with different methodological components. The process begins with generating semantic embeddings of the text, followed by dimensionality reduction and clustering of these representations to identify distinct topics. Each cluster is then linked to its most representative terms to aid interpretability. In an optional post-processing step, semantically similar topics can be merged to produce a more concise and coherent final set of topics.

We used the pre-trained all-MiniLM-L6-v2 sentence transformer[68,69] to embed the preprocessed Community Notes text into a 384-dimensional vector space. To handle the large dataset size, we first applied Principal Component Analysis (PCA)[70] for dimensionality reduction (see Table S4), followed by Uniform Manifold Approximation and Projection (UMAP)[71]. The UMAP initialisation parameters are listed in Table S5.

Clustering was conducted using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[72], which automatically estimates the number of clusters, making it well-suited for topic modelling where the number of topics is not known a priori. HDBSCAN is robust to noise and integrates effectively with UMAP-reduced data. Key hyperparameters for HDBSCAN are detailed in Table S6. We excluded a noise cluster from the list of topics. For the overall model parameters, see Table S7.

Finally, class-based Term Frequency–Inverse Document Frequency (cTF-IDF) was applied to extract interpretable topic representations from the HDBSCAN clusters.

Interactions

To examine Contributor interaction dynamics within Community Notes, we constructed monthly networks based on the Ratings.

For the study period, January 2021 to January 2025, we created 49 monthly rating files, each used to generate a corresponding interaction network.

Each network is represented as a directed graph, where nodes correspond to Community

Notes Contributors. A directed edge from a rater to a Note writer indicates that the rater evaluated the writer's Note. Each edge contains three attributes—helpful, unhelpful, and somewhat helpful—representing the number of each respective rating the rater assigned to that Note writer during the month. These complete monthly graphs, referred to as the "whole networks", are available for download.

For further analysis, each monthly network was decomposed into three subgraphs, each isolating a different interaction type: a positive interaction subgraph (edges with helpful ratings), a negative interaction subgraph (edges with unhelpful ratings), and a neutral interaction subgraph (edges with somewhat helpful ratings). In each subgraph, edge weights represent the frequency of the corresponding interaction type within that month.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Taha Yasseri (taha.yasseri@tcd.ie).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The data and code used in this study are publicly available at `https://doi.org/10.5281/zenodo.16761304`.

ACKNOWLEDGMENTS

This publication has emanated from research supported in part by grants from Taighde Éireann – Research Ireland under Grant numbers 18/CRT/6049 and IRCLA/2022/3217. TY acknowledges

support from Workday, Inc. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

AUTHOR CONTRIBUTIONS

Literature review, S.M., K.S., H.C., M.D., S.G., and T.Y.; data analysis, S.M., N.C., H.D., K.S.; drafting the paper, S.M., N.C., H.D., T.Y., supervision, T.Y; All authors approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During the preparation of this work, the authors used ChatGPT 4.0 to improve the writing style of this article. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

SUPPLEMENTAL INFORMATION INDEX

Document S1. Tables S1-S7 and their captions in the PDF.

Document S2. Table S8. Full list of topics and their representative words sorted by topic size is available at `https://doi.org/10.5281/zenodo.16761304.` .

References

1. Thai, M.T., Wu, W., and Xiong, H. (2016). Big data in complex and social networks. CRC Press.

2. Hassan, N., Adair, B., Hamilton, J.T., Li, C., Tremayne, M., Yang, J., and Yu, C. (2015). The quest to automate fact-checking. In Proceedings of the 2015 computation+ journalism symposium. Citeseer.

3. Binns, R., Veale, M., Van Kleek, M., and Shadbolt, N. (2017). Like trainer, like bot? inheritance of bias in algorithmic content moderation. In Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9. Springer pp. 405–415.

4. X Community Notes (n.d.). Signing up. https://communitynotes.x.com/guide/en/contributing/signing-up. . Accessed: 2025-09-29.

5. X Community Notes (n.d.). Notes shown on x. https://communitynotes.x.com/guide/en/contributing/notes-on-twitter. . Accessed: 2025-09-29.

6. X Community Notes (n.d.). Rating and writing impact. https://communitynotes.x.com/guide/en/contributing/writing-and-rating-impact. . Accessed: 2025-09-29.

7. X Community Notes (n.d.). Locking and unlocking the ability to write notes. https://communitynotes.x.com/guide/en/contributing/writing-ability. . Accessed: 2025-09-29.

8. X Community Notes (n.d.). Request a note. https://communitynotes.x.com/guide/en/under-the-hood/note-requests. . Accessed: 2025-09-29.

9. Yasseri, T., and Menczer, F. (2023). Can crowdsourcing rescue the social marketplace of ideas? Communications of the ACM *66*, 42–45.

10. Allen, J., Martel, C., and Rand, D.G. (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 Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–19.

11. Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M., Coleman, K., and Baxter, J. (2022). Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. arXiv preprint arXiv:2210.15723.

12. X Corp. / Community Notes Guide (n.d.). Note ranking code – under the hood (community notes guide). https://communitynotes.x.com/guide/en/under-the-hood/note-ranking-code. . Accessed: 2025-08-05.

13. Saeed, M., Traub, N., Nicolas, M., Demartini, G., and Papotti, P. (2022). Crowdsourced fact-checking at twitter: how does the crowd compare with experts? In Proceedings of the 31st ACM international conference on information & knowledge management. pp. 1736–1746.

14. Martel, C., Allen, J., Pennycook, G., and Rand, D.G. (2024). Crowds can effectively identify misinformation at scale. Perspectives on Psychological Science *19*, 477–488.

15. Pilarski, M., Solovev, K.O., and Pröllochs, N. (2024). Community notes vs. snoping: how the crowd selects fact-checking targets on social media. In Proceedings of the International AAAI Conference on Web and Social Media vol. 18. pp. 1262–1275.

16. Zhang, S., Wang, L., Shi, D., Chuai, Y., Chen, J., Chen, Y., Wang, Y., Wang, Y., Yi, X., and Li, H. (2025). Commenotes: Synthesizing organic comments to support community-based fact-checking. arXiv preprint arXiv:2509.11052.

17. Borenstein, N., Warren, G., Elliott, D., and Augenstein, I. (2025). Can community notes replace professional fact-checkers? arXiv preprint arXiv:2502.14132.

18. Renault, T., Amariles, D.R., and Troussel, A. (2024). Collaboratively adding context to social media posts reduces the sharing of false news. arXiv preprint arXiv:2404.02803.

19. Chuai, Y., Pilarski, M., Lenzini, G., and Pröllochs, N. (2024). Community notes reduce the spread of misleading posts on x. OSF Preprint. https://osf. io/preprints/osf/3a4fe, preprint: not peer reviewed.

20. Slaughter, I., Peytavin, A., Ugander, J., and Saveski, M. (2025). Community notes moderate engagement with and diffusion of false information online. arXiv preprint arXiv:2502.13322.

21. Kankham, S., and Hou, J.R. (2024). Community notes vs. related articles: Assessing real-world integrated counter-rumor features in response to different rumor types on social media. International Journal of Human–Computer Interaction pp. 1–15.

22. Drolsbach, C.P., Solovev, K., and Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. PNAS nexus *3*, pgae217.

23. Chuai, Y., Tian, H., Pröllochs, N., and Lenzini, G. (2023). The roll-out of community notes did not reduce engagement with misinformation on twitter. arXiv e-prints pp. arXiv–2307.

24. Bak-Coleman, J. (2023). Limiting factors in the effectiveness of crowd-sourced labeling for combating misinformation.

25. De, S., Bakker, M.A., Baxter, J., and Saveski, M. (2024). Supernotes: Driving consensus in crowd-sourced fact-checking. arXiv preprint arXiv:2411.06116.

26. Center for Countering Digital Hate (2024). Rated not helpful; how x's community notes system falls short on misleading election claims. . URL: `https://counterhate.com/research/rated-not-helpful-x-community-notes/` new research by CCDH shows that X's Community Notes are failing to counter false and misleading claims about the US election.

27. Bouchaud, P., and Ramaciotti, P. (2025). Algorithmic resolution of crowd-sourced moderation on x in polarized settings across countries. arXiv preprint arXiv:2506.15168.

28. Wirtschafter, V., and Majumder, S. (2023). Future challenges for online, crowdsourced content moderation: evidence from twitter's community notes. Journal of Online Trust and Safety *2*.

29. Borwankar, S., Zheng, J., and Kannan, K.N. (2022). Democratization of misinformation monitoring: The impact of twitter's birdwatch program. Available at SSRN 4236756.

30. Pröllochs, N. (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.

31. Phillips, S., Wang, S.Y.N., Carley, K.M., Rand, D., and Pennycook, G. (2024). Emotional language reduces belief in false claims. OSF pp. jn23a.

32. Simpson, K. (2022). " Obama never said that": Evaluating fact-checks for topical consistency and quality. University of Washington.

33. Solovev, K., and Pröllochs, N. (2025). References to unbiased sources increase the helpfulness of community fact-checks. arXiv preprint arXiv:2503.10560.

34. Kangur, U., Chakraborty, R., and Sharma, R. (2024). Who checks the checkers? exploring source credibility in twitter's community notes. arXiv preprint arXiv:2406.12444.

35. Chuai, Y., Zhang, S., Wang, Z., Yi, X., Mosleh, M., and Lenzini, G. (2025). Request a note: How the request function shapes x's community notes system. arXiv preprint arXiv:2509.09956.

36. Fraxanet, E., Pellert, M., Schweighofer, S., Gómez, V., and Garcia, D. (2024). Unpacking polarization: Antagonism and alignment in signed networks of online interaction. PNAS nexus *3*, pgae276.

37. Champaigne, R. (2022). Birdwatch and the polarization of the crowds.

38. Kim, J., Wang, Z., Shi, H., Ling, H.K., and Evans, J. (2025). Differential impact from individual versus collective misinformation tagging on the diversity of twitter (x) information engagement and mobility. Nature Communications *16*, 973.

39. Megill, C., Barry, E., and Small, C. (2022). "coherent mode" for the world's public square. arXiv preprint arXiv:2211.12571.

40. Renault, T., Mosleh, M., and Rand, D.G. (2025). Republicans are flagged more often than democrats for sharing misinformation on x's community notes. Proceedings of the National Academy of Sciences *122*, e2502053122.

41. Benjamin, G. (2021). Who watches the birdwatchers? sociotechnical vulnerabilities in twitter's content contextualisation. In International Workshop on Socio-Technical Aspects in Security. Springer pp. 3–23.

42. Wang, C., and Lucas, P. (2024). Efficiency of community-based content moderation mechanisms: A discussion focused on birdwatch. Group Decision and Negotiation *33*, 673–709.

43. Augenstein, I., Bakker, M., Chakraborty, T., Corney, D., Ferrara, E., Gurevych, I., Hale, S., Hovy, E., Ji, H., Larraz, I. et al. (2025). Community moderation and the new epistemology of fact checking on social media. arXiv preprint arXiv:2505.20067.

44. Lloyd, T., Nguyen, T., Levy, K., and Naaman, M. (2025). Beyond community notes: A framework for understanding and building crowdsourced context systems. arXiv preprint arXiv:2509.15434.

45. Mujumdar, R., and Kumar, S. (2021). Hawkeye: a robust reputation system for community-based counter-misinformation. In Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 188–192.

46. Li, H., De, S., Revel, M., Haupt, A., Miller, B., Coleman, K., Baxter, J., Saveski, M., and Bakker, M.A. (2025). Scaling human judgment in community notes with llms. arXiv preprint arXiv:2506.24118.

47. Costabile, L., Orlando, G.M., La Gatta, V., and Moscato, V. (2025). Assessing the potential of generative agents in crowdsourced fact-checking. arXiv preprint arXiv:2504.19940.

48. Juncosa Calahorrano, M.G., Mohammadi, S., Samahita, M., and Yasseri, T. Interplay between diversity and efficiency in collaborative efforts for content moderation (2025). Forthcoming.

49. Mohammadi, S., and Yasseri, T. (2025). Ai feedback enhances community-based content moderation through engagement with counterarguments. arXiv preprint arXiv:2507.08110.

50. Ortega, F., Gonzalez-Barahona, J.M., and Robles, G. (2008). On the inequality of contributions to wikipedia. In Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008). IEEE pp. 304–304.

51. Yasseri, T., and Kertész, J. (2013). Value production in a collaborative environment: socio-physical studies of wikipedia. Journal of Statistical Physics *151*, 414–439.

52. Sauermann, H., and Franzoni, C. (2015). Crowd science user contribution patterns and their implications. Proceedings of the national academy of sciences *112*, 679–684.

53. Glenski, M., Pennycuff, C., and Weninger, T. (2017). Consumers and curators: Browsing and voting patterns on reddit. IEEE Transactions on Computational Social Systems *4*, 196–206.

54. Nguyen, D., McGillivray, B., and Yasseri, T. (2018). Emo, love and god: making sense of urban dictionary, a crowd-sourced online dictionary. Royal Society open science *5*, 172320.

55. Ad Fontes Media, Inc. (2025). Ad fontes media. https://adfontesmedia.com/. . Accessed: 2025-08-05.

56. Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment *2008*, P10008.

57. Loomba, S., De Figueiredo, A., Piatek, S.J., De Graaf, K., and Larson, H.J. (2021). Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. Nature human behaviour *5*, 337–348.

58. Enders, A.M., Uscinski, J., Klofstad, C., and Stoler, J. (2022). On the relationship between conspiracy theory beliefs, misinformation, and vaccine hesitancy. Plos one *17*, e0276082.

59. Allcott, H., and Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of economic perspectives *31*, 211–236.

60. Greene, C.M., Nash, R.A., and Murphy, G. (2021). Misremembering brexit: Partisan bias and individual predictors of false memories for fake news stories among brexit voters. Memory *29*, 587–604.

61. Klein, O. (2025). Mobilising the mob: The multifaceted role of social media in the january 6th us capitol attack. Javnost-The Public *32*, 33–50.

62. Kaplan, J. (2025). More speech and fewer mistakes. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/. . Accessed: (February 25, 2025).

63. TikTok Newsroom (2025). Footnotes – tiktok newsroom. https://newsroom.tiktok.com/en-us/footnotes. . Accessed: 2025-08-05.

64. Truong, B.T., Kim, S., Nogara, G., Verdolotti, E., Sahneh, E.S., Saurwein, F., Just, N., Luceri, L., Giordano, S., and Menczer, F. (2025). Delayed takedown of illegal content on social media makes moderation ineffective. arXiv preprint arXiv:2502.08841.

65. X Corp. / Community Notes Guide (n.d.). Download data – under the hood (community notes guide). https://communitynotes.x.com/guide/en/under-thehood/downloaddata.. Accessed: 2025-08-05.

66. Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

67. Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794.

68. Reimers, N., and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

69. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in neural information processing systems *33*, 5776–5788.

70. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: Machine learning in python. the Journal of machine Learning research *12*, 2825–2830.

71. McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

72. McInnes, L., Healy, J., Astels, S. et al. (2017). hdbscan: Hierarchical density based clustering. J. Open Source Softw. *2*, 205.

SUPPLEMENTAL INFORMATION

Additional Tables

Table S1: URL Extraction Regex Pattern

Component	Matches
`(?:https?|ftp)://§+`	Protocol-based URLs (HTTP/HTTPS/FTP)
`www\.§+`	`www`-prefixed domains
`[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}(/§*)?`	Domain strings with TLDs and optional paths

Table S2: Domain Extraction Regex Pattern

Pattern	**Behavior**
`(?:https?://)?`	Optional protocol
`(?:www\.)?`	Optional `www` prefix
`([^/]+)`	Captures text until first / (core domain)

Table S3: Domain Normalisation Rules

Original	**Normalised**
`twitter.com`, `x.com`, `t.co`	`x.com`
`youtu.be`	`youtube.com`
`bbc.co.uk`	`bbc.com`
`*.wikipedia.org`	`wikipedia.org`

Table S4: Changes to default PCA hyperparameters in scikit-learn (v1.2.2)

Parameter	Value
n_components	100
random_state	42

Table S5: Modified hyperparameters for the UMAP model (v0.5.7) applied to PCA-reduced embeddings.

Parameter	Value
n_components	10
n_neighbors	30
min_dist	0.1
random_state	42

Table S6: Modified hyperparameters for HDBSCAN (v0.8.40) used to cluster the UMAP embeddings.

Parameter	Value
min_cluster_size	500
min_samples	20
cluster_selection_epsilon	0.15
prediction_data	False

Table S7: BERTopic (v0.16.4) initialisation using pre-computed UMAP embeddings and HDBSCAN clusters, with passthrough functions specified for the embedding and clustering models.

Parameter	Value
embedding_model	None
umap_model	PassthroughUMAP()
hdbscan_model	passthrough_hdbscan
nr_topics	60
verbose	True
low_memory	True