Can Crowdsourcing Rescue the Social Marketplace of Ideas?

How collaborative designs and community-based moderation could improve social media.

In the late 1980s, Tim Berners-Lee invented the World Wide Web to facilitate information sharing and collaboration among researchers. Rapid and widespread adoption by academic and commercial entities followed in the 1990s. After the burst of the dot-com bubble at the dawn of the new millennium, several social Web platforms deeply transformed the information ecosystem within a few years. Wikis, blogs, social tagging and sharing sites, online social networks, and microblogs all significantly lowered the costs of producing and sharing content. This transition was hailed as a harbinger of more democratic participation in the information and knowledge economy—the realization of the “marketplace of ideas.” No project was a better icon of such utopian dreams than Wikipedia: the revolutionary vision of an encyclopedia that can be read, and more importantly, edited by anyone, surprised media scholars and practitioners. The potential of the largest online collaborative project in history was soon realized, and the quality of Wikipedia articles reached a level comparable to professionally edited encyclopedias despite its reported limitations and biases. Facebook and Twitter appeared a few years later. Ten years into the millennium, democratic revolutions were credited to these social media platforms and phenomena such as the Arab Spring were referred to as the Facebook revolution.

It did not take long for this sweet story to take a bitter twist. Hate speech, misinformation, polarization, manipulation, extremism, and data privacy scandals became commonplace online. In January 2021, the spread of disinformation and conspiracy theories on social media led to one of the darkest moments in the history of U.S. democracy, when the U.S. Capitol was attacked on January 6. In the same month, Wikipedia celebrated its 20th birthday, enduring the test of time as a reputable source of information. How is it possible that one crowd-based social Web technology stands as the most successful example of collaborative and healthy information sharing, while the other is blamed for epistemic chaos? What can this para-
dox teach us as we strive to fix the broken marketplace of ideas?

Both Facebook and Twitter have announced community-based review platforms to address misinformation. Can these crowdsourcing approaches work on social media? We provide an overview of the potential affordances of community-based approaches to content moderation based on past research on collaborative systems such as Wikipedia and preliminary analysis of data from Twitter. While our analysis generally supports a community-based approach to content moderation, it also warns against potential pitfalls, particularly when the implementation of the new infrastructure focuses on crowd-based “validation” rather than “collaboration.”

To be sure, there are many important differences between Wikipedia and social media platforms, including design elements, user motivations and characteristics. The business model of social media, with its emphasis on engagement over quality, is also contrasted with the consensus-based model of Wikipedia. However, a review of past research points to the network effects of content generation and diffusion as a key to understanding how community-based moderation could rescue the social media marketplace of ideas, provided there is a serious intention by the commercial platforms to promote a healthier information environment.

The Collaborative Content Generation Model
There are several successful community-based information-sharing endeavors. Citizen science projects, where nonexpert volunteers collaboratively contribute to scientific discoveries through community interaction, show that such models can be robust.11 Wikipedia is arguably the most successful community-based information-sharing initiative and for sure the most researched. Despite its generally acknowledged reliability, any experienced Wikipedia editor can tell stories of “edit wars” in which they have found themselves. Wikipedia edit wars have been studied in detail, showing in most cases they lead to a consensus after enough time.24 Agent-based simulations have helped reveal the secret of consensus-reaching mechanisms in Wikipedia. Articles, as shared products of editorial activities, facilitate indirect interactions between editors with opposing opinions, who would not engage in any positive interactions otherwise.22 These indirect interactions provide both mechanisms and incentives to bring editors closer to one another and reach a consensus. Even though the simulated models are calibrated and validated against empirical data from Wikipedia, at an abstract theoretical level they suggest that any crowd-based system can reach a consensus if there is a collaborative element, regardless of the level of existing polarization. The same models were used to examine the effects of banning editors with extreme opinions. Counterintuitively, the exclusion of extremists hinders the consensus-reaching process.16 Moreover, a moderate amount of controversy is correlated with higher article quality17 and higher completeness and neutrality ratings. Wikipedia is not free of bias and abuse, such as inauthentic and coordinated accounts that manipulate content. Over time, the platform has dealt with these threats by evolving a complex set of community rules and norms along with a hierarchy of user and administrator roles, privileges, and tools (including benign bots23).

The Homophily-Based Model
As opposed to Wikipedia, where articles are meant for a diverse community, the content created and shared on social media is targeted at an echo-chamber structure in the friend/follower network of a sample of news-sharing Twitter users (a). Node color is based on partisanship, inferred from the political alignment of shared news sources. Node size is proportional to the fraction of shared links to low-credibility sources. (Source: Nikolov, D. et al.) The network structure of the 1,500 most active Birdwatch users and their positive ratings of each other’s notes (b). Node colors are determined by a community detection algorithm and node size indicates the number of interactions.

Both Facebook and Twitter have announced community-based review platforms to address misinformation.
have reported that social networks on Twitter\(^4\) and Facebook\(^19\) tend to be structured around ideologically homogeneous echo chambers (see the accompanying figure) and that users embedded in partisan echo chambers are exposed to more low-credibility information.\(^14\) Homophily, the tendency to connect with like-minded individuals, leads to homogeneous communities both online and offline. However, social media mechanisms make it very easy to create as well as to cut off a social tie. Platforms encourage us to “block” users whose content we dislike. Mathematical analysis\(^4\) and agent-based simulations\(^18\) show that such costless unfollowing accelerates the formation of segregated echo chambers. Influence and social pressure ensure strong homogeneity of opinions within an echo chamber. One can generate engagement and gain popularity more easily by addressing a homogeneous audience than by appealing to common ground across diverse communities.

Another important ingredient of social media is algorithmic bias. Newsfeed ranking algorithms favor content that is likely to trigger engagement. Popularity also plays a role in Wikipedia, where content tends to be generated around topics that attract attention.\(^3\) But on social media, engagement is directly amplified by the platforms as it is a crucial element in their business model. This leads to greater exposure to low-quality content\(^6\) as well as ideologically aligned content that confirms or reinforces existing beliefs.\(^13\) Algorithmic bias also increases vulnerability to manipulation by inauthentic or coordinated accounts. The incentives created by all of these social media characteristics—targeting homogeneous audiences as well as amplification of low-quality, confirmatory, and manipulated content—go against reaching a consensus outside of one’s group. Worse, hate speech and verbal abuse are natural byproducts of this highly polarized information ecosystem and can spill over to real-world violence.\(^9\)

**Community-Based Solutions**

Ronald Reagan once said: “Peace is not the absence of conflict, it is the ability to handle conflict by peaceful means.” The current social media approach of hiding conflict through self-selection, unfollowing, removal of content, and account bans neither prevents nor mitigates conflict. Some researchers suggested exposing users to counter-attitudinal content, a positive algorithmic bias designed to break the bubbles. However, experiments show that partisan users become more entrenched in their beliefs once they are exposed to opposing views.\(^3\) Moreover, mathematical models suggest that no matter how much positive algorithmic bias is applied, as long as unfollowing is costless, echo chambers will inevitably form.\(^4\) What is effective, in contrast, is sharing personal experience\(^12\) and—if we have learned one thing from the Wikipedia experience—collaborative interaction. Such collaboration in the context of social media could be aimed at tackling misinformation and community policy violations. Recent experiments suggest that crowd-sourced layperson judgments can be effective at identifying misinformation.\(^15\) Such a community approach could scale up fact-checking and moderation practices while mitigating both misinformation and polarization.\(^1\)

Following this line of argument, Facebook announced a community review program\(^21\) in December 2019 and Twitter launched a community platform to address misinformation\(^7\) in January 2021. Here, we focus on Twitter's platform, originally called Birdwatch and later renamed Community Notes by Elon Musk, for which some preliminary data is available. In the current implementation, a member of the group of reviewers (selected by Twitter based on undisclosed criteria) can add a note to a tweet they find “misinformed or potentially misleading.” A note provides some information selected from predefined values about the tweet (misleading factual error, misleading satire) as well as some free text where the reviewer can comment and link to external sources. Then other reviewers express their agreement or disagreement with the existing notes through additional annotations such as helpfulness and informativeness. Ultimately, notes produced by reviewers will become visible next to the corresponding tweets based on the support/opposition they have received from other reviewers.

We analyzed Birdwatch helpfulness ratings as of December 2022—586,462 ratings of 31,422 notes by 15,023 reviewers. We observed evidence of a highly polarized network with two clusters where reviewers agree with those in the same group and disagree with those in the opposite group. To quantify polarization among reviewers, we considered pairs of reviewers who rated each other at least once. Of these pairs of reviewers with reciprocal ratings, 60%–70% are consistent in that both rate each other as helpful (in the same cluster) or not helpful (in different clusters). Furthermore, we looked at triads of reviewers with reciprocal ratings to quantify structural balance, which occurs when all three reviewers agree with each other (in the same cluster) or when two reviewers agree with each other (in the same cluster) and disagree with the third (in the other cluster). Even though 39% of ratings are negative, we found that only 25% of triads are structurally imbalanced. This indicates a high level of cohesion within the clusters and disagreement across the clusters when interactions exist. The second part (b) of the accompanying figure offers visual confirmation of these findings by mapping the network of Birdwatch reviewers. An edge between two nodes represents reciprocal ratings that indicate agreement on average. The polarization among Birdwatch reviewers mimics the one observed among generic Twitter users (see part (a) of the figure here).
The design of community review systems will have to be modified to enforce collaboration rather than competitive behavior.

In particular, while crowdsourcing initiatives may present viable options to rescue the marketplace, collaborative designs are crucial; direct collaboration seems to be a key element in successful crowd-based systems.

Technological advancements in Web-supported collaboration are not yet fully exploited in the area of content moderation and a dire need for innovative designs is apparent. The unit of content analysis could be more than a single tweet or post and a more holistic, conversation-based consensus building could be the goal of the content moderators. The quality of conversations can be determined and enhanced by a diverse set of moderators and this may facilitate a quality-based promotion system instead of the current ones that are mainly based on engagement. In addition to providing a better-functioning content moderation system, these features will hinder astroturfing and other tools of social media-based propaganda.

References
1. Allen, J. et al. Scoping up fact-checking using the wisdom of crowds. Science Advances 7, 36 (2021); DOI:10.1126/sciadv.abj4353
2. Allen, J. et al. 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 CHI Conference on Human Factors in Computing Systems (2022); DOI:10.1145/3491102.3502040
3. Bai, C.A. et al. Exposure to opposing views on social media can increase political polarization. In Proceedings of the National Academy of Sciences 115, 37 (2018); DOI:10.1073/pnas.1804840115
4. Bles, C. and Yasseri, T. Positive algorithmic bias cannot stop fragmentation in homophilic networks. The Journal of Mathematical Sociology 46, 1 (2022); DOI:10.1080/0022250X.2020.1818076
5. Campaglia, G.L. et al. The production of information in the attention economy. Scientific Reports 5, 1 (2015); DOI:10.1038/srep09452
6. Campaglia, G.L. et al. How algorithmic popularity bias hinders or promotes quality. Scientific Reports 8, 1 (2018); DOI:10.1038/s41598-018-34203-2
7. Coleman, K. Introducing Birdwatch, a community-based approach to misinformation. Twitter (2021); https://web.archive.org/web/20211011030341/https://twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation
8. Conover, M.D. et al. Partisan asymmetries in online political activity. EPU Data Science 1, 6 (2012); DOI:10.31140/epjds/4
9. Gattacker, J.D. et al. Online engagement between opposing political protest groups via social media is linked to physical violence of opposing encounters. Social Media & Society 7, 1 (2021); DOI:10.1177/2053951719886445
10. Giles, J. Internet encyclopedia go head to head by Nature 439, 7070 (2005); DOI:10.1038/388190a
11. Ibrahim, K. et al. Gender imbalance and spatiotemporal patterns of contributions to citizen science projects: The case of Zooniverse. Frontiers in Physics 8, 650720 (2021); DOI:10.3389/fphy.2021.650720
12. Kubin, E. et al. Personal experiences bridge moral extremism divides better than facts. In Proceedings of the National Academy of Sciences 118, 6 (2021); DOI:10.1073/pnas.2008389118
13. Levy, R. Social media, news consumption, and polarization: Evidence from a field experiment. American Economic Review 111, 3 (2021); DOI:10.1257/aer.20191177
14. Miklov, D. et al. Right, red, and left, partisanship predicts (asymmetric) vulnerability to misinformation. HKS Misinformation Review 1, 7 (2021); DOI:10.37016/frr-2020-55
15. Pennycook, G. and Rand, D.G. Fighting misinformation on social media using crowdsourced judgments of news source quality. In Proceedings of the National Academy of Sciences 116, 7 (2019); DOI:10.1073/pnas.1808781116
16. Rubis, C. et al. Understanding and coping with extremism in an online collaborative environment: A data-driven model. PLoS ONE 12, 3 (2017); DOI:10.1371/journal.pone.0173561
17. Ruplec, T. et al. Relating Wikipedia article quality to edit behavior and link structure. Applied Network Science 5, 61 (2020); DOI:10.1007/s41109-020-00305-y
18. Sasahara, K. et al. Social influence and unfollowing accelerate the emergence of echo chambers. Journal of Computational Social Science 4, 1 (2020); DOI:10.1007/s42001-020-00084-7
19. Schmidt, A.L. et al. An analysis of news consumption on Facebook. Proceedings of the National Academy of Sciences 114, 12 (2017); DOI:10.1073/pnas.1617052114
20. Shi, P. et al. The wisdom of polarized crowds. Nature Human Behaviour 3, 4 (2019); 329–336; DOI:10.1038/s41562-019-0154-6
21. Silverman, H. Helping fact-checkers identify false claims faster. Facebook (2019); https://web.archive.org/web/20210929132905/https://about.fb.com/news/2018/12/helping-fact-checkers/
22. Torkis, J. et al. Opinions, conflicts, and consensus: Modeling social dynamics in a collaborative environment. Physical Review Letters 110, 8 (2013); DOI:10.1103/physrevlett.110.088703
23. Tsvetkova, M. et al. Even good bots fight: The case of Wikipedia. PLoS ONE 12, 7 (2017); DOI:10.1371/journal.pone.0171717
24. Yasseri, T. et al. Dynamics of conflicts in Wikipedia. PLoS ONE 7, 6 (2012); DOI:10.1371/journal.pone.0038869

Taha Yasseri (taha.yasseri@ucd.ie) is an associate professor of Sociology and a George Fellow in Public Policy at University College Dublin, Dublin, Ireland.

Filippo Menczer (fil@indiana.edu) is the Luddy Distinguished Professor of Informatics and Computer Science and the director of the Observatory on Social Media at Indiana University, Bloomington, IN, USA.

We are grateful to David G. Rand, Janos Kertesz, Ksenia Musaeelyan, and Dimitar Nikolov for their feedback and comments on the manuscript. Taha Yasseri is partially supported by Google. Filippo Menczer is supported in part by Knight Foundation and Craig Newmark Philanthropies. The opinions and recommendations expressed here are those of the authors and do not necessarily reflect the opinion of the funders. The authors declare no competing interests.

Copyright held by authors.