Can crowdsourcing rescue the social marketplace of ideas?

Taha Yasseri [1], Fil Menczer [2]

[1]School of Sociology, University College Dublin, Dublin, Ireland
[2]Observatory on Social Media, Indiana University, Bloomington, USA

***Keywords**: Homophily, Misinformation, Crowdsourcing, Community, Social Media*

How is it possible that some social web technology such as Wikipedia stand as the most successful example of collaborative and healthy information sharing, while the others, such as Twitter are blamed for epistemic chaos? To be sure, there are many important differences between Wikipedia and social media platforms, including design elements, business models, and user motivations and characteristics. However, a review of past research points to the network effects of content generation 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 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 [1]. What is effective, in contrast, is sharing personal experience 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 crowdsourced layperson judgments can be effective at identifying misinformation [2]. Such a community approach could scale up fact-checking and moderation practices while mitigating both misinformation and polarization.

Following this line of argument, Facebook announced a community review program in December 2019 and Twitter launched a community platform to address misinformation4 in January 2021. Here we focus on Twitter's platform, called Birdwatch, for which some preliminary data is available. In the current Birdwatch implementation, a member of the group of reviewers (selected by Twitter based on undisclosed criteria) can add a note to a tweet that 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 February 2022 — 189,744 ratings of 17,888 notes by 7,884 reviewers. We observed evidence of a highly balanced network with two well-separated clusters where reviewers agree with those in the same group and disagree with those in the opposite group. In fact, of the pairs of reviewers with reciprocal ratings, 71% are consistent in that they both rate each other helpful (in the same cluster) or not helpful (in different clusters). Furthermore,

despite 35% negative ratings, we found that only 22% of triads of reviewers with reciprocal ratings are structurally imbalanced, i.e., inconsistent with all three reviewers agreeing with each other (in the same cluster) or with two reviewers in agreement with each other (in the same cluster) and in disagreement with the third (in the other cluster). Fig. 1 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.

It is unlikely that this polarization is a reflection of objective arguments; rather, it merely represents the political affiliations of the reviewers. Analysis of the notes confirms that users systematically reject content from those with whom they disagree politically [3]. One might argue that the population of Birdwatch reviewers is less homogenous than that of Wikipedia editors. While this may be true, a polarized crowd can be even more effective in producing high-quality content compared with a homogenous team [4]. The missing ingredient, however, is collaboration: reviewers of opposing opinions currently do not have to reach a consensus. The design of Birdwatch will have to be modified to enforce collaboration rather than competitive behaviour; robustness to competition is as critical as resistance to coordinated manipulation. Wikipedia teaches us that community rules can enforce such norms.

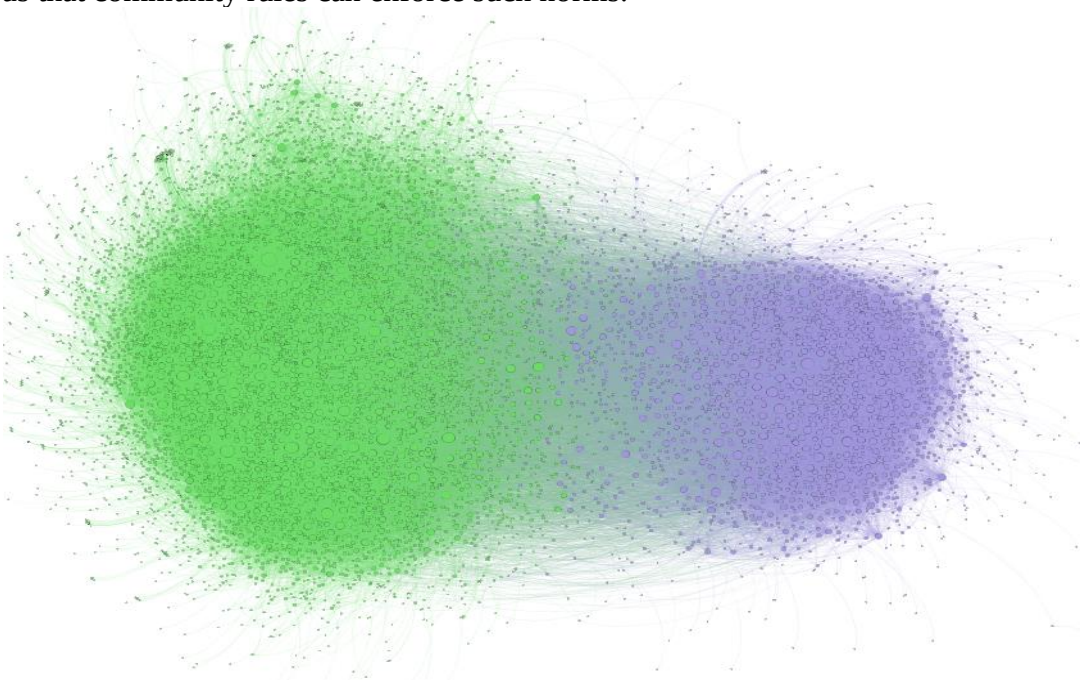

Fig. 1: The network structure of Birdwatch users and their positive ratings of each other's notes. Node colours are determined by a community detection algorithm and node size indicates the number of interactions.

[1] Kubin, E. et al. 2021. Personal experiences bridge moral and political divides better than facts. *PNAS*. 118, 6 (Jan. 2021), e2008389118.
[2] Nikolov, D. et al. 2021. Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. *HKS Misinformation Review*. 1, 7 (Feb 2021).
[3] Allen, Jennifer et al. 2021. Birds of a Feather Don't Fact-check Each Other. Preprint PsyArXiv.
[4] Shi, F. et al. 2019. The wisdom of polarized crowds. Nature Human Behaviour. 3, 4 (Mar. 2019), 329–336.