Learning and enforcing a cultural consensus in online communities

Necdet Gürkan (ngurkan@stevens.edu)
School of Business, Stevens Institute of Technology
Hoboken, NJ 07030, USA

Jordan W. Suchow (jws@stevens.edu)
School of Business, Stevens Institute of Technology
Hoboken, NJ 07030, USA

Abstract

Online communities rely on their members to understand and follow community norms, which they learn by observing others and the consequences of their behavior, seeing codes of conduct, and receiving feedback via moderation. Here, to determine the contribution of each source of learning to the preservation of a social norm, we extend cultural consensus theory, a mathematical framework for identifying the cultural consensus in a community. In particular, we extend the model to include learning from experience, centralized moderation, and decentralized moderation, three features commonly found in online communities. We then apply the extended model to data from an online community dedicated to preserving a norm related to the psychophysical scaling of intersubjective notions of beauty derived from facial aesthetics. We find that users' perceptual alignment with the norm before enculturation predicts involvement in the community and that experience in the community is an important indicator for group perceptual learning.

Keywords: cultural consensus, Bayesian modeling, online communities, face perception

Introduction

Social interactions are influenced by social norms that determine what actions are considered acceptable and unacceptable (Cialdini & Goldstein, 2004). Social norms shape human behavior and perception in many domains, including beauty judgments (Sugiyama, 2005; Bergstrom & Neighbors, 2006), facial expressions (Hareli, Kafetsios, & Hess, 2015), economic decision making (Azar, 2004), and health (Staunton, Louis, Smith, Terry, & McDonald, 2014). Social norms are important to social conduct and have impacts that are extensive, effective, enduring, and frequently unnoticed (McDonald & Crandall, 2015). Social norms are theorized to help people coordinate (Gelfand, Harrington, & Jackson, 2017), cooperate (Fehr & Schurtenberger, 2018), avoid social sanctions (Wanders, Homan, van Vianen, Rahal, & Van Kleef, 2021), and earn rewards (Klucharev, Hytönen, Rijpkema, Smidts, & Fernández, 2009).

Though communities are often more successful when everyone follows the social norms (Kimbrough & Vostroknutov, 2016; Fehr & Schurtenberger, 2018), norms are not always followed or even understood by everyone in the community. To mitigate the possible detrimental consequences of disobedience, cohesive groups are more likely to enforce group norms via social sanctioning when someone does not comply with the social norms (Horne, 2007). Enforcement of norms can promote desirable behaviors and suppress undesirable behaviors in the community.

It is often difficult to measure social norms and their effects because of the complexity and heterogeneity of norm-driven behaviors. Some of the prominent and widely used methods to measure them are incentive-compatible elicitation tasks (Krupka & Weber, 2013), the reasoned action approach (Fishbein & Ajzen, 2011), and the normative and empirical expectation approach (Bicchieri, 2016). These measurement methods have been limited to laboratory games and self-report studies. However, many forms of social interactions are best studied in the environment in which they occur in order to fully understand the factors that influence human behavior (Parigi, Santana, & Cook, 2017).

The increasing number of social interactions that take place in virtual spaces provide a unique opportunity to study human behavior, one that has necessitated the innovative adaptation of methods that have previously been used for studying the "real world" (Hine, 2000). Many methods have been developed to study the cultural and social norms of online communities and their downstream effects on in-person behavior (Kozinets, 2010; Wilson & Peterson, 2002; De Souza & Preece, 2004). The observation of naturally occurring and fully observable interactions in online communities can help to answer the question of how social norms are formed, learned, and enforced.

Cultural consensus theory (CCT) is a statistical framework that can be used to infer what cultural beliefs influence social practices and the degree to which individuals know or show those beliefs (Romney, Weller, & Batchelder, 1986). These models provide an opportunity to study individual differences in whether a member of a group understands the consensus answer to a question among a community and allows people to differ in both their level of cultural knowledge and response biases. Researchers have applied the CCT framework to find a practical and concise definition of beliefs that are accepted by a group that shares common knowledge. These models have been widely used to study mental health (Alang, 2018), cognitive evaluation (Heshmati et al., 2019), eye-witness testimony (Waubert de Puiseau, Aßfalg, Erdfelder, & Bernstein, 2012), and organizational culture (Rinne & Fairweather, 2012). For example, Oravecz and Vandekerckhove (2020) combined a cultural consensus model and a dynamical model into a single joint process model to exam-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

ine whether subjective beliefs of what makes people loved are connected to daily life experience of love.

Here, we extend the cultural consensus model to include three important features commonly found in online communities: individual learning, centralized moderation, and decentralized moderation. We then apply the extended model to study an online community dedicated to preserving a norm related to the psychophysical scaling of intersubjective notions of beauty derived from facial aesthetics. The model is able identify the cultural consensus, each member's competence, individual differences in the rate of learning the social norm, and the effects of centralized and decentralized moderation.

The plan of the paper is as follows. We begin by reviewing moderation in online communities and the CCT model for continuous responses. Next, we describe the online community studied here and introduce our extension to the CCT model. Finally, we present and discuss the results of fitting the model to behavior observed in the online community.

Moderation in online communities

Online (or internet) communities are groups of individuals with a shared interest or purpose who use the internet to communicate with each other. Online communities have their own sets of guidelines, norms, and needs, such as moderation, engagement, and management (Kraut & Resnick, 2011). To thrive, online communities must have active exchanges of information that improves the community's objectives, for example, regular commenting on a discussion post (McWilliam, 2000).

One form of moderation technique is a centralized moderation, which is performed by one or more members who have the designated role of *moderator*. Moderators' roles are to keep the activities courteous and beneficial. Grimmelmann (2015) argued that the success of online communities depend heavily on the behavior of moderators. Moderators have the ability to promote or conceal postings, and to recruit or prohibit users to maintain social norms within the community. Because a moderator's strict governing can cause attrition and disengagement (McWilliam, 2000), moderators must carefully cultivate precise norms to promote appropriate member activity in online communities.

Another form of moderation is decentralized moderation, where community members provide feedback to other members' actions in the form of comments, likes, dislikes, upvotes, downvotes, and similar symbolic forms of reaction. Reading threads is often a primary method by which individuals acquire the direct informational and social benefits available from a community (Welser, Gleave, Fisher, & Smith, 2007). Individuals who provide feedback contribute new information resources that help others (Lakhani & Von Hippel, 2004). Those who engage in decentralized moderation work to actively maintain and promote actions in the community by guiding discussions towards collectively agreed norms (Lampe & Resnick, 2004).

Active moderation helps to build vigorous shared norms among a community's members. Therefore, having a rough consensus about the community norms can enhance engagement and guide community purposes (Kiesler, Kraut, Resnick, & Kittur, 2012). For instance, Wikipedia built a robust neutral-point-of-view norm that encouraged members of the community to write a trustworthy encyclopedia (Wikipedia, 2021a); likewise, the norm where editors take additional precautions when adding information on living people can lessen the risk of a lawsuit (Wikipedia, 2021b)

Every online community must incorporate successive generations of newcomers to survive. However, newcomers often engage in behaviors that are considered to be a violation of the community's norms. During early interactions with newcomers before enculturation, the community must protect itself from potentially harmful behaviors that may arise when dealing with newcomers (Kraut, Burke, Riedl, & Resnick, 2012). In order to overcome these issues and have more committed newcomers, moderators and members can help newcomers to learn the norms of the community (Preece & Shneiderman, 2009).

Cultural consensus theory

Cultural consensus theory (CCT) is a mathematical framework that measures each respondent's cultural knowledge while estimating the culturally "right" answers to a series of questions defined by group beliefs or norms. The CCT model jointly estimates (1) an individual's level of cultural knowledge from their agreement with a consensus truth and (2) the consensus truth itself from a weighted average of responses, giving higher weight to individuals with higher cultural competence. The first CCT model, the General Condorcet Model (GCM), was developed for binary data (true/false responses) and assumes that the consensus truth of each item is also a binary value. The GCM has been widely used in the social and behavioral sciences (Weller, 2007).

Batchelder and Anders (2012) introduced an alternate assumption to the GCM to extend it to continuous truths. An extensive CCT model for ordinal data was developed using a Gaussian appraisal model (Anders & Batchelder, 2015). In addition, CCT models for continuous response data were developed to estimate and detect cultural consensuses, informant knowledge, response biases, and item difficulty from continuous data (Anders, Oravecz, & Batchelder, 2014; Batchelder & Romney, 1988; Batchelder, Strashny, & Romney, 2010).

Here, we describe the Continuous Response Model (CRM), developed in Anders et al. (2014) to allow for multiple consensus truths; however, we will describe the single-consensus-truth version of the model, which serves as the basis for our extension to the model.

The single-truth CRM

Data fit by the CRM consists of observations of the random response profile matrix $\mathbf{X}_{ik} = (X_{ik})_{N\times M}$ for N respondents and M items, where each respondent's response falls within $(0,1)$ or a finite range that permits a linear transformation to $(0,1)$.

The CRM links the random response variables in (0, 1) to the real line with the logit transform, $X^* = \text{logit}(X_{ik})$. Therefore, each item also has a consensus value in $(-\infty, \infty)$.

The single-truth CRM is specified by the following five axioms, which were developed to model the continuous response of respondents that differ in competency, E_i, and response biases, a_i and b_i, to items that have different shared latent truth values. The respondents have a latent appraisal of these item values with a mean at the item's consensus location plus some error, which depends on their competence level and the item difficulty. Axiom 1 locates the item truth values in the continuum. Axiom 2 defines the appraisal error is normally distributed with mean zero. Axiom 3 sets the appraisals are conditionally independent given the respondents' cultural truth and the error standard deviations. Axiom 4 specifies the standard appraisal error that depends on the respondent's competence and item difficulty. Axiom 5 covers each respondent's response bias and location response tendencies on the scale.

Axiom 1. Cultural Truth. There is a single consensus truth

$$\mathbf{Z}^* = (Z_k^*)_{1\times M}, \text{where each } Z_k^* \in (-\infty, \infty). \tag{1}$$

Axiom 2. Latent Appraisals. Each respondent draws a latent appraisal for each item $Y_{ik} = Z_k^* + \varepsilon_{ik}$. The ε_{ik} error variables are distributed normally with mean 0 and standard deviation σ_{ik}.

Axiom 3. Conditional Independence. The ε_{ik} are mutually stochastically independent.

Axiom 4. Precision. There are knowledge competency parameters $\mathbf{E} = (E_i)_{1\times N}$ with all $E_i > 0$, and item difficulty parameters specific to each cultural truth $\Lambda = (\lambda_k)_{1\times M}, \lambda_k > 0$ such that

$$\sigma_{ik} = \lambda_k / E_i. \tag{2}$$

If all item difficulties are equal, then each λ_k is set to 1.

Axiom 5. Response Bias. There are two respondent bias parameters that act on each respondent's latent appraisals, Y_{ik}, to arrive at the observed responses, the X_{ik}. These include a scaling bias, $\mathbf{A} = (a_i)_{1\times N}, a_i > 0$; and shifting bias $\mathbf{B} = (b_i)_{1\times N}, -\infty < b_i < \infty$, where

$$X_{ik}^* = a_i Y_{ik} + b_i. \tag{3}$$

These five axioms undergird the single-truth CRM that the present work will extend.

Extending the cultural consensus model

Here, we extend the cultural consensus model to include learning and moderation, both centralized and decentralized. Our computational model is designed to represent how a user contributes to a community and how the community responds to those contributions. We use a Bayesian hierarchical model that allows for multiple processes to contribute to a single set of observed data (Lee, 2011; Anders et al., 2014). The hierarchical structure of our generative model of ratings and moderation events is described below.

The extended cultural consensus model assumes that each posting k has a consensus response (e.g., rating or relevance):

$$Z_k \sim \text{Beta}(\alpha, \beta),$$

where α and β both sampled from a prior distribution, $\alpha, \beta \sim \text{Gamma}(10, 1)$. We assume that each post has the same level of difficulty in responding to it.

Learning

In the extended model, each user has a cultural competence that determines the precision with which they can access the cultural consensus. Critically, competence is assumed to improve as a function of experience in a process of exponential saturation, such that for user i,

$$E_i(t) = E_i(0) + (E_\infty - E_i(0))(1 - e^{-L_i t}), \tag{4}$$

where t is the number of ratings previously completed by the user, $E_i(0)$ is the user's initial competence, E_∞ is the asymptotic competence under infinite practice, assumed to be shared across all users, and L_i is a user-specific rate parameter. Users may vary in their initial competence due to incidental alignment to the cultural norm or because they were lurkers who observed the community and learned from it, but did not participate for some time. Observers may learn from observing other group members' behavior and from obtaining feedback from other users in the form of centralized and decentralized moderation.

A user's initial cultural competence is determined by their knowledge about the cultural consensus before enculturation and is sampled as follows:

$$E_i(0) \sim \text{Gamma}(\kappa_a, \lambda_a),$$

where κ_a and λ_a are both sampled from a prior distribution, $\kappa_a, \lambda_a \sim \text{Exponential}(0.4)$.

Each user has a learning rate that determines how quickly their competence approaches its asymptotic value with experience and is sampled from

$$L_i \sim \text{Gamma}(\kappa_b, \lambda_b),$$

where κ_b and λ_b are both sampled from a prior distribution $\kappa_b \sim \text{Exponential}(1)$ and $\lambda_b \sim \text{Exponential}(0.001)$. In this application, these hyperparameter choices provide the model with considerable flexibility to express individual differences in learning the social norm.

We assume that all users share an asymptotic competence, E_∞, the highest attainable level of competence and is sampled by

$$E_\infty \sim \text{Gamma}(\kappa_a, \lambda_a),$$

where κ_a and λ_a are sampled from prior distributions $\kappa_a \sim \text{Exponential}(0.4)$ and $\lambda_a \sim \text{Exponential}(0.4)$, respectively.

Each rating is then assumed to be generated from

$$b_i \sim \text{Normal}(0, 0.1)$$
$$r \sim \text{Normal}(Z_k + b_i, 1/E_i(t)), \tag{5}$$

where b_i is a user-level bias term. We assume that the scaling bias for each user is 1. Each rating, r, is sampled from the

normal distribution with the mean of item consensus, Z_k, plus the user shifting bias, b_i, and inverse variance, given by the competence $E_i(t)$.

Decentralized moderation

Incorporating decentralized moderation enables us to obtain more precise estimates of cultural competence and consensus by considering the way that the community responds to the individual. Different communities implement decentralized moderation in different ways. For example, Reddit associates a score attribute with each post and comment. The score is the difference between the number of upvotes and downvotes for that post or comment. Because Reddit does not provide direct access to the proportion of votes that are upvotes, nor the total number of voters, these are latent parameters of the model. Our generative model of a score s associated with a response r begins by sampling the number of voters n, then proceeds to sample the proportion φ of votes that are upvotes under the assumption that upvotes fall away exponentially with distance from the consensus according to a rate parameter Δ:

$$\begin{aligned} n &\sim \text{Exponential}(0.4) \\ \Delta &\sim \text{Uniform}(0,100) \\ \varphi &= e^{-\Delta|Z_k - r|} \\ s &= n\varphi - n(1-\varphi) \\ &= n(2\varphi - 1). \end{aligned} \tag{6}$$

The distance is calculated as the difference between the response and the consensus.The score is derived from the total number of voters and the proportion of voters that upvote the response.

Centralized moderation

Incorporating centralized moderation also enables more precise estimates of cultural competence and consensus. Moderators are responsible for enforcing community guidelines and norms to support the community's goals. In online communities, moderators actively inspect posts for relevance. We assume all moderators have high-precision estimates of the cultural consensus. If the response is not aligned with the consensus, moderators will moderate the comment along with a warning. However, moderators are not perfectly vigilant: they notice misaligned responses with probability less than one. Our generative model of moderation events begins by sampling a vigilance probability

$$g \sim \text{Uniform}(0,1)$$

We further assume that moderators have a limited sensitivity τ to error with respect to the consensus when moderating:

$$\tau \sim \text{Uniform}(0,5).$$

We use the vigilance probability and the moderators' sensitivity as an input to a sigmoidal link function to cast warnings:

$$y = \frac{2g}{1 + e^{-\tau(r - Z_k)}} - g \tag{7}$$

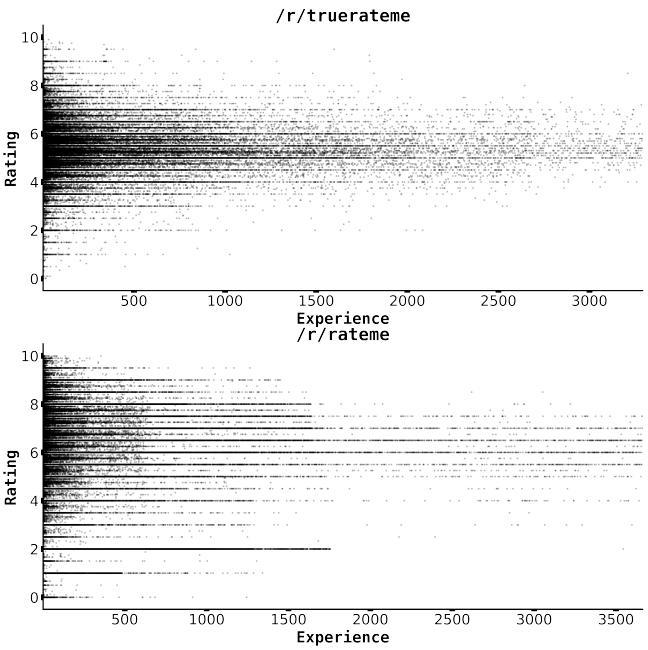

Figure 1: Ratings vary with experience (user's ordinal position of the rating). The upper and lower plots represent the /r/truerateme and /r/rateme communities, respectively.

The link function produces output in the interval $[-1,1]$. We flip a coin weighted $|y|$ to determine whether a moderation event happens, and if so, the sign of y governs the direction of moderation.

Data description

Reddit is a social media platform where users share posts and comment on them. Reddit is divided into millions of *subreddits*, communities that cover a variety of subjects (e.g., r/television, r/askscience, and r/movies). Users can upvote or downvote posts and comments on posts, which can in part determine who else sees the posts and comments.

Here, we study the r/truerateme and r/rateme communities, which are dedicated to rating attractiveness based on facial aesthetics in response to face images posted by their members. The rating scales are between 0 (lowest rating) to 10 (highest rating). Studying these communities thus provides a unique opportunity to study intersubjective cultural phenomena because of how amenable these particular norms are to quantitative characterization. Though these communities have similar purposes, they adopt and enforce different social norms. In r/truerateme, the community preserves a particular intersubjective beauty norm by defining a rating scale and using strict moderation, which is not the case in the /rateme community, which focuses more on subjective impressions of attractiveness.

We curated a dataset with all comments and voting behavior in these two communities from the Pushshift Reddit dataset (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, 2020). For each dataset, we considered all submissions

made between October 1, 2017, and June 30, 2021. Because the /r/truerateme dataset contains outdated upvote and downvote data, we refreshed it using the Reddit API. We extracted all numerical ratings from user comments and selected the midpoint of ratings that took the form of ranges (e.g. 5.5–6). After filtering the comments to include only top-level comments with ratings, the datasets for r/truerateme and r/rateme contained 173,331 and 119,105 ratings, respectively.

We use only the dataset from the /r/truerateme community to build our computational model because this community enforces strict norms that provide a strong reason to believe that there is a cultural consensus to be estimated. The data consist of an ordered set of ratings by users to given items, as well as the scores (decentralized moderation) and moderation events (centralized moderation) associated with the rating. The response tensor for ratings is given by

$$\mathbf{X} = (X_{ikn})_{16{,}718\times 46{,}107\times 3{,}293},$$

where i is the user index, k is the item index, and n is the user's *experience*, the number of responses that a user has thus far contributed to the community. We rescale ratings to fall in the interval $[0,1]$. Note that the ratings tensor is sparse because members rate only a subset of items. Next, there is the score matrix, given by

$$\mathbf{D} = (D_{ikr})_{16{,}718\times 46{,}107\times 173{,}331},$$

where i is the user index, k is the item index and r is the given rating. Cells contain scores, which are calculated as the number of upvotes minus the number of downvotes associated with the rating.

And finally, there is the moderation matrix, given by

$$\mathbf{H} = (H_{ikr})_{16{,}718\times 46{,}107\times 173{,}331},$$

where i is the user index, k is the item index and r is the given rating. Cells can take one of three values: a label for an underrating moderation, an overrating moderation, or no moderation.

Implementation: The model was implemented in NumPyro (Phan, Pradhan, & Jankowiak, 2019) with the JAX backend (Bradbury et al., 2020). The model components were integrated into a single likelihood function and a set of prior distributions, needed to infer a posterior over the unobserved variables in our model using the No-U-Turn Sampler (NUTS) (Hoffman, Gelman, et al., 2014), a standard Markov chain Monte Carlo sampling algorithm, as implemented in Numpyro. We used 1 chain with 2,000 warm up samples and 2,000 draw samples, thereby obtaining 2,000 posterior samples. We ensured that the posterior had converged by ensuring there were not divergence transitions.

Results

Enforcing the norm: Enforced norms establish the identity of a group that enhances the differences between group members and those from outside of the group. The two online communities studied here, r/truerateme and r/rateme, enforce different norms of facial attractiveness (Fig. 1). The

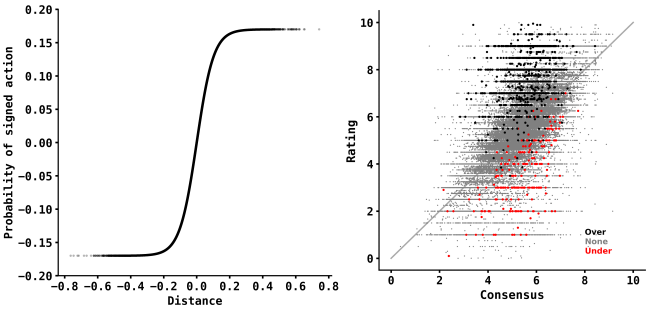

Figure 2: **Left**: Inferred shape of the function linking error to the probability of the moderator commenting that a rating was an overrating vs. underrating. **Right**: Calibration curve showing when moderators comment.

/r/rateme community permits a larger range of ratings compared to the /r/truerateme community, and ratings do not converge towards any particular group agreement despite time spent in the community (Fig. 1).

The moderators in /r/truerateme are sensitive to misalignment to the consensus (Fig. 2). Upper and lower asymptotes, corresponding to the best-fit vigilance parameter, 0.17, show that moderators often do not comment on even large errors (Fig. 2, left). Community members show similar sensitivity when providing feedback to other members' ratings (Fig. 3). These results support the notion that /r/truerateme strictly enforces its inter-subjective social norms to maintain their community purposes.

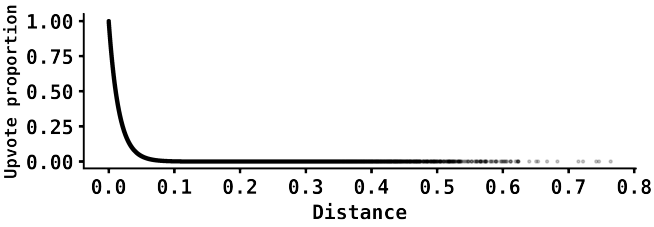

Figure 3: The proportion of voters who upvote decays with the distance between the rating and the consensus value. The best-fit rate parameter, displayer here, was 65.19.

Learning the norm: Starting when a user first joins a community, the venues for learning multiply. Newcomers can start to learn about community norms even before they interact with other community members. Social learning theory by Bandura (1978) suggests that individuals learn by observing how others behave. As shown in Fig. 4, we find that users' perceptual alignment with the community norm before enculturation predicts involvement in the community ($r = 0.38$). Also, newcomers' initial observations and interactions might influence their involvement in the community. Competent newcomers in rating face images are more likely to receive positive feedback and satisfy their expectations by choosing the right community.

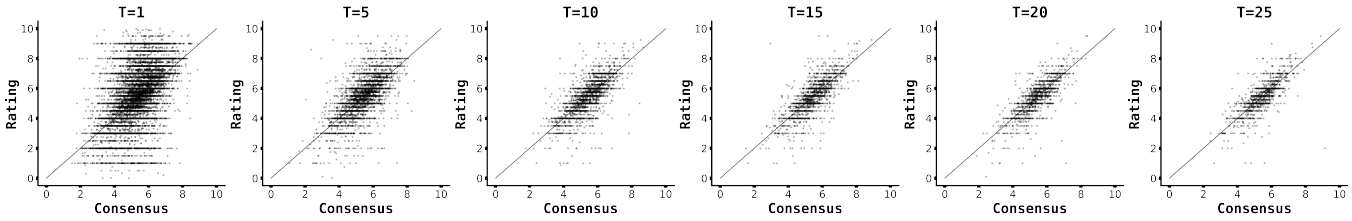

Figure 6: Users with more experience are better calibrated to the consensus.

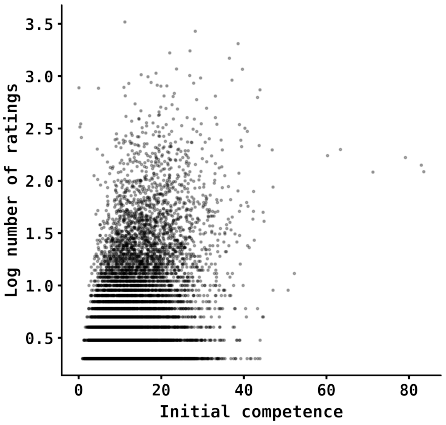

Figure 4: Tenure in the community is predicted by initial competence.

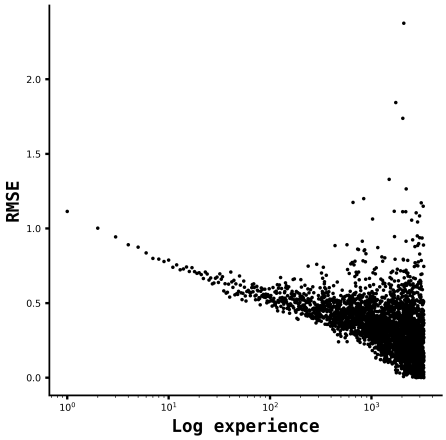

Figure 5: Experience helps the group perceptual alignment.

Fig. 5 shows that experience in rating items is an important indicator for group learning of the shared social norms ($r = -0.55$). As the members are maturing, the group learns the norms of the community both through observation and direct reinforcement, where accurate ratings are rewarded while inaccurate ratings are punished. Once the members reach a certain maturity, the consensus becomes more available to them.

Although the /r/truerateme community demonstrates quick perceptual adaptation after the first rating in the community to recalibrate their responses to a shared social norm (Fig. 6), it is unclear whether the primary mechanism that supports group learning is individual perceptual learning or sanctioning and dropout of community members with low competence.

Discussion

We built a computational model that extends Cultural Consensus Theory to include (1) learning, (2) centralized moderation, and (3) decentralized moderation, all commonly observed mechanisms in online communities. We study the contribution of each social interaction to the learning and enforcement of social norms in an online community.

This study considers a particular online community, /r/truerateme, that provides a unique opportunity to leverage psychophysical methods to understand intersubjective cultural phenomena because of how amenable this community norm is to quantitative characterization. We note that our model is applicable to study a broad array of intersubjective judging processes within finite scales that are formed by social norms.

Communities have various ways to sanction acceptable and unacceptable social behaviors in the society. In most of the online communities, members can up vote/down vote or like/dislike the behavior to display their agreement with comments and posts made by other members. We leverage this observed data in our hierarchical model to determine item consensus, moderation errors, and user competence more precisely.

The evaluation of intersubjective judgements has been assessed based on quantitative evidence, such as a evaluation of judges in sport competitions (Heiniger & Mercier, 2018). Likewise, moderators are allowed to judge members' behaviors that are based intersubjective norms to control group behavior. However, moderators can have biases and low competence in judging these behaviors. To quantitatively assess moderator performance, future work can extend our computational model to infer each moderator's competence along with their biases in support of providing more transparent mechanisms for promoting the health of online and other communities.

References

Alang, S. (2018). Contrasting depression among african americans and major depressive disorder in the DSM-V. *Journal of Public Mental Health*, *17*, 11-19.

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*(1), 151–181.

Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, *61*, 1–13.

Azar, O. H. (2004). What sustains social norms and how they evolve?: The case of tipping. *Journal of Economic Behavior & Organization*, *54*(1), 49–64.

Bandura, A. (1978). Social learning theory of aggression. *Journal of communication*, *28*(3), 12–29.

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*(5), 316–332.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*(1), 71–92.

Batchelder, W. H., Strashny, A., & Romney, A. K. (2010). Cultural consensus theory: Aggregating continuous responses in a finite interval. In *International Conference on Social Computing, Behavioral Modeling, and Prediction* (pp. 98–107).

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 830–839).

Bergstrom, R. L., & Neighbors, C. (2006). Body image disturbance and the social norms approach: An integrative review of the literature. *Journal of Social and Clinical Psychology*, *25*(9), 975–1000.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., & Wanderman-Milne, S. (2020). JAX: composable transformations of Python+ NumPy programs, 2018. *URL http://github. com/google/jax*, *4*, 16.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621.

De Souza, C. S., & Preece, J. (2004). A framework for analyzing and understanding online communities. *Interacting with Computers*, *16*(3), 579–610.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458–468.

Fishbein, M., & Ajzen, I. (2011). *Predicting and changing behavior: The reasoned action approach*. Psychology press.

Gelfand, M. J., Harrington, J. R., & Jackson, J. C. (2017). The strength of social norms across human groups. *Perspectives on Psychological Science*, *12*(5), 800–809.

Grimmelmann, J. (2015). The virtues of moderation. *Yale JL & Tech.*, *17*, 42-69.

Hareli, S., Kafetsios, K., & Hess, U. (2015). A cross-cultural study on emotion expression and the learning of social norms. *Frontiers in Psychology*, *6*, 1501.

Heiniger, S., & Mercier, H. (2018). Judging the judges: A general framework for evaluating the performance of international sports judges. *arXiv preprint arXiv:1807.10055*.

Heshmati, S., Oravecz, Z., Pressman, S., Batchelder, W. H., Muth, C., & Vandekerckhove, J. (2019). What does it mean to feel loved: Cultural consensus and individual differences in felt love. *Journal of Social and Personal Relationships*, *36*(1), 214–243.

Hine, C. (2000). The virtual objects of ethnography. *Virtual ethnography*.

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623.

Horne, C. (2007). Explaining norm enforcement. *Rationality and Society*, *19*(2), 139–170.

Kiesler, S., Kraut, R., Resnick, P., & Kittur, A. (2012). Regulating behavior in online communities. *Building Successful Online Communities: Evidence-based Social Design*, 125–178.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, *14*(3), 608–638.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140–151.

Kozinets, R. V. (2010). *Netnography: Doing ethnographic research online*. Sage publications.

Kraut, R. E., Burke, M., Riedl, J., & Resnick, P. (2012). The challenges of dealing with newcomers. *Building Successful Online Communities: Evidence-based Social Design*, 179–230.

Kraut, R. E., & Resnick, P. (2011). Encouraging contribution to online communities. *Building Successful Online Communities: Evidence-based Social Design*, 21–76.

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524.

Lakhani, K. R., & Von Hippel, E. (2004). How open source software works:"free" user-to-user assistance. In *Produktentwicklung mit virtuellen Communities* (pp. 303–339). Springer.

Lampe, C., & Resnick, P. (2004). Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 543–550).

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7.

McDonald, R. I., & Crandall, C. S. (2015). Social norms and social influence. *Current Opinion in Behavioral Sciences*, *3*, 147–151.

McWilliam, G. (2000). Building stronger brands through online communities. *MIT Sloan Management Review*, *41*(3), 43-54.

Oravecz, Z., & Vandekerckhove, J. (2020). A joint process model of consensus and longitudinal dynamics. *Journal of Mathematical Psychology*, *98*, 102386.

Parigi, P., Santana, J. J., & Cook, K. S. (2017). Online field experiments: studying social interactions in context. *Social Psychology Quarterly*, *80*(1), 1–19.

Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv preprint arXiv:1912.11554*.

Preece, J., & Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-computer Interaction*, *1*(1), 13–32.

Rinne, T., & Fairweather, J. (2012). A mixed methods approach: Using cultural modeling and consensus analysis to better understand new zealand's international innovation performance. *Journal of Mixed Methods Research*, *6*(3), 166-183.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*(2), 313–338.

Staunton, M., Louis, W. R., Smith, J. R., Terry, D. J., & McDonald, R. I. (2014). How negative descriptive norms for healthy eating undermine the effects of positive injunctive norms. *Journal of Applied Social Psychology*, *44*(4), 319–330.

Sugiyama, L. S. (2005). Physical attractiveness in adaptationist perspective. *In D. M. Buss (Ed.), The Handbook of Evolutionary Psychology*, 292–343.

Wanders, F., Homan, A. C., van Vianen, A. E., Rahal, R.-M., & Van Kleef, G. A. (2021). How norm violators rise and fall in the eyes of others: The role of sanctions. *PLoS one*, *16*(7), e0254574.

Waubert de Puiseau, B., Aßfalg, A., Erdfelder, E., & Bernstein, D. M. (2012). Extracting the truth from conflicting eyewitness reports: A formal modeling approach. *Journal of Experimental Psychology: Applied*, *18*(4), 390.

Weller, S. C. (2007). Cultural consensus theory: Applications and frequently asked questions. *Field methods*, *19*(4), 339–368.

Welser, H. T., Gleave, E., Fisher, D., & Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, *8*(2), 1–32.

Wikipedia. (2021a). *Neutral point of view.* Retrieved from `https://en.wikipedia.org/wiki/Wikipedia:Neutral\_point\_of\_view`

Wikipedia. (2021b). *Sockpuppetry.* Retrieved from `https://en.wikipedia.org/wiki/Wikipedia:Sockpuppetry`

Wilson, S. M., & Peterson, L. C. (2002). The anthropology of online communities. *Annual review of Anthropology*, *31*(1), 449–467.