The Effects of Feedback on Human Behavior in Social Media: An Inverse Reinforcement Learning Model

Sanmay Das
Washington University in St. Louis
sanmay@seas.wustl.edu

Allen Lavoie
Washington University in St. Louis
allenlavoie@wustl.edu

ABSTRACT

We introduce and validate a learning model of human behavior change in response to feedback on social media. People who participate in these types of websites, like Wikipedia, Reddit, and others, are learning agents whose choices about how to allocate their effort are dynamic and responsive to how they feel their efforts were received in the past. By explicitly taking into account the reinforcement effects of different types of feedback received on prior contributions, our model is able to significantly outperform all known baselines in predicting future contributions both on synthetic data and on real data collected from the social news site reddit.com. Our model has an intuitive interpretation as users playing mixed strategies in a game-like setting with thousands of other users and thousands of available pure strategies. In this interpretation, our task is then *inverse reinforcement learning*: recovering users' reward functions based on observed behavior.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence

Keywords

Social Media; Social Simulation; Multi-Agent Learning

1. INTRODUCTION

Information sharing is an increasingly social phenomenon. Websites like Facebook and Twitter allow users to share links and snippets of text with networks of other users. More structured social news sites such as Slashdot, Digg, and Reddit focus on public discussions of recent news. The popularity of these venues has raised new questions concerning the quality of information and societal effects of democratized information dissemination. However, participants are no longer passive consumers: information sharing is inherently social. What effect does this sharing have on the sharer, and what can this tell us about the dynamics of social media?

Effects on user behavior of social-psychological feedback have been documented recently in social media on YouTube and Digg [25], and on Wikipedia [26]. While social news is often viewed as a way for participants to influence public awareness and opinion, the act of sharing and its associated social feedback have a much more direct effect on those doing the sharing. What are the implications of millions of users providing and receiving feedback, influencing and being influenced? We liken social media to a game, where a user's strategy helps to determine the social feedback received by others, and the choices made by other users influence a user's own utility. As users learn and adapt their strategies, they create and abandon groups, communities, and whole venues. Understanding the complex social dynamics governing the evolution of these communities is a key challenge for those who study multiagent systems and collective intelligence. In this paper, we investigate how our interests are determined, individually and collectively, by feedback from others.

To motivate the problem, consider the example in Figure 1. It shows the influence of feedback in determining future effort allocation on Reddit. Reddit is a social news website where participants can vote and comment on links/items posted by others. It is partitioned into special interest communities, or subreddits, and users spend their time on some subset of these subreddits. The figure shows the relative change in effort spent by a user on a subreddit as a function of the number of replies received to that user's comment in the subreddit. Getting more replies to a comment leads to a user being much more likely to spend more time on that subreddit in the future. It is clear that understanding social-psychological community dynamics well enough to model them depends crucially on understanding the role of feedback—like responses to comments—in incentivizing future effort by individual participants.

We introduce a model of behavior in response to social-psychological feedback in social media. This model builds on work in human game playing with matrix games [3], in the behavioral/ reinforcement learning tradition. Rather than attempting to find an optimal strategy, players make updates to mixed strategies in response to the feedback they receive. We combine this learning model with a more sophisticated model of initial preferences (based on the Hierarchical Dirichlet Process [23]), and create an inference algorithm which discovers the dynamics of the learning process itself along with factors which constitute behavior-altering feedback.

We test the algorithm on real and synthetic social media data, where users choose between thousands of communities based on their initial preferences and the feedback they receive. On synthetic data, the inference algorithm is able to recover a user's true distribution of community preferences—a mixed strategy under the game analogy—with near perfect accuracy. We then apply the learning model to real data from the social news site Reddit, which at any one time is composed of thousands of active communities. The learning model outperforms a plethora of static and adaptive baselines on this probabilistic prediction task. Moreover, the model provides easily interpretable explanations. It allows for a form of inverse reinforcement learning where probabilistic human behavior

Appears in: *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), Lomuscio, Scerri, Bazzan, Huhns (eds.), May, 5–9, 2014, Paris, France.*

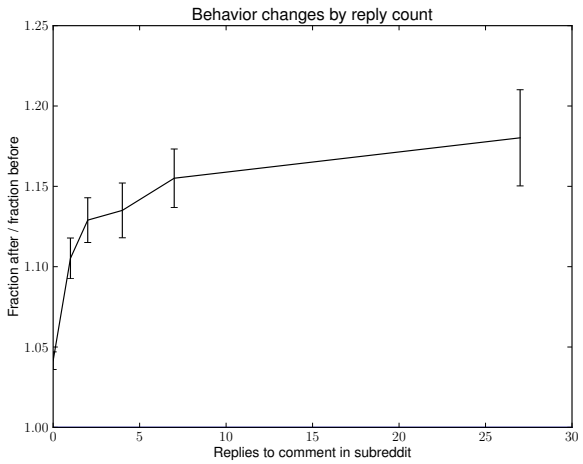

Figure 1: The ratio of the fraction of time spent on a subreddit after a comment to the fraction of time spent on that subreddit before the comment (excluding the contribution itself in both cases), as a function of the number of replies to that contribution. Receiving more responses makes a participant more likely to allocate more of their effort to that subreddit in the future, consistent with a learning effect in response to social feedback.

changes are viewed as the product of a set of features, whose relative importance can then be determined by the inference algorithm.

1.1 Related work

There has been significant interest in the population-level dynamics of collective intelligence. Wu and Huberman [24] study the relationship between a novelty factor of content and its popularity. Szabo and Huberman [22] predict the long-term popularity of content given the initial reaction to it. Networks play an important part in social media dynamics. Lerman and Hogg [15, 14] introduce stochastic models of popularity which distinguish network effects and visibility effects resulting from the content venue when predicting and explaining content popularity. Lerman *et al* [16] show that social proximity predicts sharing behavior in social media. Heaukulani and Ghahramani [8] present a Bayesian model for reasoning about unobserved interactions in social networks. We begin with a more basic model of individuals, where the existence of interactions are more important than the identities of the other participants in those interactions.

Another line of research considers the effects of social interactions and social biases in social media. Wu *et al* [25] find in social media that users who contribute have an increased propensity to contribute in the future, and that this effect along with social interactions explain the distribution of the number of contributions per user. Cobot [11] was developed to track and respond to social feedback in a virtual world. On Wikipedia, Zhu *et al* [26] test the effects of positive and negative feedback on users' propensities to contribute in the future. Huberman *et al* [10] show that attention predicts future contributions on YouTube, and that a lack of attention is demotivating. Muchnik *et al* [18] study the effects of social influence bias by conducting a randomized experiment where vote counts are seeded with a small positive or negative value. They find asymmetric herding effects as a result of these initial signals. Hsieh *et al* [9] study predictors of volunteer socialization on Reddit. Gilbert [6] finds that users are unlikely to participate in evaluating new content on Reddit, citing a lack of social interaction as one potential reason. To the best of our knowledge, we are the first to quantify the effects of social feedback on individual user behavior by explicitly modeling complex future decisions.

We make an analogy between social feedback and the rewards received in a game-theoretic setting. There is related work studying social media in game-theoretic terms, and studying human behavior in game playing. Munie and Shoham [19] put wikis, ratings and similar collaborative venues into a game theoretic framework. They show that users taking actions under a simple myopic rule converge to an equilibrium. Genter *et al* [5] investigate ways to control groups of agents exhibiting flocking behavior, reminiscent of our community seeding experiments. Erev and Roth [3] study a model of human game playing, explaining the changes in strategies resulting from the rewards received in repeated matrix games. They show that simple learning models outperform equilibrium predictions which assume rational behavior. We adapt this model to explain and predict the effects of social feedback in social media.

Inverse reinforcement learning [20] and apprenticeship learning [1] recover the utility functions of actors based on their actions, with the hope producing similar (often optimal) policies in unseen situations. Knox and Stone [12, 13] add manual signals in a reinforcement learning setting, allowing humans to affect policies and speed learning. Chernova and Veloso [2] deal with the stochasticity of human demonstrations of policies in an inverse reinforcement learning setting using Gaussian mixture models. We are agnostic as to whether the behavior we are learning about is truly optimal: we wish to learn how behavior changes in response to the results of prior actions, by learning utility functions and update rules from observed behavior that are predictive of future behavior.

2. MODEL

Our goal is to model the behavior of users on social media sites such as Reddit. One feature of interest on these websites is strong community structure. On Reddit, for example, users choose to belong to communities called subreddits, which are user-run and organized around a theme: anything from New York City to cat pictures. What drives community selection? We make an analogy to games: users choose to post in a specific community, analogous to picking a pure strategy or action. Based on the action, they get a reward in the form of social feedback from other users who have also chosen that community. Users play this repeated community selection game, giving and receiving social feedback. Rather than explicitly specifying a goal, we focus on jointly learning how players adapt to rewards and what form those rewards take.

2.1 Background

We begin with some background from previous work, then present a model for players in repeated community selection games.

2.1.1 Human game playing

We are interested in how humans play community selection games. Our model builds on the three-parameter model of Erev and Roth [3] for humans playing mixed strategies in relatively simple matrix games with small, fixed numbers of strategies. A player begins with equal propensities for playing each pure strategy k:

$$q_k(1) = Z$$

Propensities are updated with a non-negative reward (which can be achieved in matrix games by subtracting the minimum reward). For a reward R after playing strategy k, we have:

$$q_k(t+1) = q_k(t) + R$$

Propensities for other strategies $j \neq k$ remain unchanged ($q_j(t+1) = q_j(t)$) under the one-parameter model. When picking a strategy, a player selects from her normalized propensities:

$$p_k(t) = q_k(t) / \sum_k q_k(t)$$

The three parameter model adds recency and exploration parameters ϕ and ϵ to the initial propensity strength parameter Z. For a reward R after taking action k:

$$q_k(t+1) = (1-\phi)q_k(t) + (1-\epsilon)R$$

Other actions are also updated under the three-parameter model. For a strategy j which was not taken:

$$q_j(t+1) = (1-\phi)q_j(t) + R\epsilon/(M-1)$$

Where M is the number of available strategies.

2.1.2 Hierarchical Dirichlet Process

The strategy space of this game, the space of all communities, has several interesting properties. First, it is not finite: users are free to start their own communities at any time. It is also quite large, with thousands of active communities at any one time. Finally, users start the game with strong prior preferences over strategies: given that a user is from New York City, she has a good chance of remaining active in that community. Likewise for hobbies, interests, organizations, and so on. Further, some communities are much more popular than others across all users.

Initial propensities are not of great importance in matrix games with small and finite strategy spaces. However, in games of the kind we are considering, with infinite strategy spaces over which users may have strong prior preferences, the representation of the initial propensities becomes critical. The model must imply a proper probability distribution over this infinite strategy space, and should also be a natural model for preferences. To this end, we adapt a nonparametric Bayesian model used in machine learning and clustering, the Hierarchical Dirichlet Process [23], as a model for initial propensities; we present only the necessary special case here. This model allows global preferences, meaning that some strategies may be more popular overall across all users, and also models a user's personal prior preferences. First, an infinite discrete distribution β—representing global preferences over strategies—is drawn from an infinite-dimensional Dirichlet distribution with concentration parameter γ:

$$\beta \mid \gamma \sim \lim_{L\to\infty} \text{Dirichlet}(\gamma/L, \ldots, \gamma/L)$$

Next, second-level distributions are drawn according to the base measure β and concentration parameter α_0:

$$\pi_j \mid \alpha_0, \beta \sim \text{Dirichlet}(\alpha_0\beta)$$

Each π_j is again an infinite discrete distribution, sharing the same "atoms" as β. This distribution π_j is our model of a user's initial propensities.

2.2 Model description

We use these models of human reinforcement learning and of initial preferences to create a generative model of human behavior in response to social-psychological feedback in large social processes. Under this generative model, we first draw global preferences $\beta \sim \lim_{L\to\infty} \text{Dirichlet}(\gamma/L, \ldots, \gamma/L)$, where $\gamma \sim \text{Gamma}(\gamma_\alpha, \gamma_\beta)$ is the first-level concentration parameter. We also draw a second-level concentration parameter $\alpha_0 \sim \text{Gamma}(\alpha_{0\alpha}, \alpha_{0\beta})$, used for generating user-specific initial preferences.

Algorithm 1 Pseudo-code for the generative model of a single user's behavior. The symbol "$\leftarrow$" denotes assignment, and "$\sim$" indicates a draw from a probability distribution.

$q^0 \sim \text{Dirichlet}(\alpha_0\beta)$ ▷ Initial propensities from HDP
$q \leftarrow q^0$
for $i \in C_u$ **do** ▷ For each of this user's actions (in order)
  $s_i \sim \text{Categorical}(q/\sum_j q_j)$ ▷ Strategy picking
  $q \leftarrow q(1-\phi)$ ▷ Forgetting
  $q_{s_i} \leftarrow (1-\epsilon)R(r_i) + q_{s_i}$ ▷ Direct reward
  $q \leftarrow q + \epsilon R(r_i) q^0$ ▷ Exploration

γ, α_0, β	Hierarchical Dirichlet Process[23] parameters
ϕ	Forgetting (beta prior)
ϵ	Exploration (beta prior)
q^0	Initial propensities for a user
R	Reward function (same for all users)
C_u	Sequence of actions by user u
r_i	Feedback/reward features for action i
s_i	Strategy of action i (observed)

Table 1: Summary of notation.

Instead of fixing the global parameters of the learning model, we also sample these from their respective prior distributions: exploration $\epsilon \sim \text{Beta}(\epsilon_\alpha, \epsilon_\beta)$ and forgetting $\phi \sim \text{Beta}(\phi_\alpha, \phi_\beta)$. These parameters are analogous to those in the three-parameter model of Erev and Roth[1]. We assume that the reward function R is linear in a set of non-negative "reward features" with non-negative weights, the weights having independent Gamma priors.

Finally, users play the game-analog: repeatedly picking actions, receiving rewards, and updating their mixed strategy based on the reward received and the global learning model. Algorithm 1 describes this process formally, and Table 1 summarizes the notation. Users draw strategies from their current propensities. For each decision made, a user "forgets," scaling down his weights by $1-\phi$ and creating a recency effect. Users receive a reward $R(r_i)$, depending on the reward features r_i. An exploration parameter ϵ distributes some fraction of this reward to the user's initial propensities q^0, with the rest going to the propensity of the chosen strategy. Each user repeats this process, choosing strategies and learning based on the resulting reward.

We do not explicitly model the reward features r_i received in response to an action, aside from the aforementioned assumption of non-negativity. These features will typically rely on the actions of other players, as they do in our experiments. This implies an additional mechanism which translates from all user strategies to reward features for each user; we do not model it. In the setting of social media, this mechanism specifies who replies to whom given where users choose to make their comments. To simplify notation, we assume all rewards accrue before the user takes a new action.

3. INFERENCE

Having specified a generative model for learning in response to social-psychological feedback, our goal is to reverse this process, making inferences about user learning from observed data. The main idea will be to separate inferences about users' initial preferences from inferences about the learning process[2]. Having done

[1]The third parameter, strength of initial propensities, is redundant in our model with the coefficients of the reward features, which can be scaled to simulate any initial propensity strength.

[2]Note that this model, where contributions come from a mixture

Algorithm 2 Pseudo-code for the reinforcement learning simulation, which forms part of the inference algorithm.

$q_{\text{init}} \leftarrow 1$
$q \leftarrow \vec{0}$
for $i \in C_u$ **do** ▷ For each of this user's actions (in order)
  $q^i, q^i_{\text{init}} \leftarrow q, q_{\text{init}}$ ▷ Record current weights
  $q \leftarrow q(1-\phi)$ ▷ Forgetting
  $q_{\text{init}} \leftarrow q_{\text{init}}(1-\phi)$
  $q_{s_i} \leftarrow (1-\epsilon)R(r_i) + q_{s_i}$ ▷ Direct reward
  $q_{\text{init}} \leftarrow q_{\text{init}} + \epsilon R(r_i)$ ▷ Exploration

this, we use Gibbs sampling for approximate Bayesian inference.

To implement this separation, we begin with a series of binary latent variables, one associated with each action by a user. These variables are sampled according to their full conditional distributions given the values of all other latent variables (i.e. Gibbs sampling). That is:

$$p(\iota_i = \text{Initial} \mid s_i, \cdot) = \frac{p(s_i \mid \iota_i = \text{Initial}, \cdot)p(\iota_i = \text{Initial} \mid \cdot)}{\sum_{\iota'_i} p(s_i \mid \iota_i = \iota'_i, \cdot)p(\iota_i = \iota'_i \mid \cdot)} \tag{1}$$

Where the sum in the denominator is over the two possible assignments of ι_i: initial or reinforcement. We use $\cdot$ as shorthand for conditioning on the values of all of the latent and observed variables except those pertaining to i (s_i, ι_i, and several we have not yet introduced). This application of Bayes' rule allows us to condition on both these latent variables and the observed action s_i.

To compute the "prior" probabilities (those not relying on s_i) in Equation (1), we refer to the simulation in Algorithm 2. Having run this simulation, which relies on the values of the latent learning parameters, we can explicitly compute those probabilities as follows:

$$p(\iota_i = \text{Initial} \mid \cdot) = \frac{q^i_{\text{init}}}{q^i_{\text{init}} + \sum_j q^i_j} \tag{2}$$

$p(\iota_i = \text{Reinforcement} \mid \cdot)$ is simply $1 - p(\iota_i = \text{Initial} \mid \cdot)$. This leaves the probability of observing strategy s_i given the assignment of ι_i. In the case that $\iota_i = \text{Initial}$, this is the probability of drawing s_i from the Hierarchical Dirichlet Process conditioned on all of the other initial strategy observations (but not any of those where $\iota_j = \text{Reinforcement}$). We reproduce this probability here, but see Teh *et al* [23] for background and details:

$$p(s_i \mid \iota_i = \text{Initial}, \cdot) = \frac{\alpha_0 \beta_{s_i} + n^{-i}_{u_i, s_i}}{\alpha_0 + n^{-i}_{u_i}} \tag{3}$$

$$n^{-i}_{u_i, s_i} = \sum_{j \in C_{u_i} \setminus i} I(\iota_j = \text{Initial and } s_j = s_i)$$

$$n^{-i}_{u_i} = \sum_{j \in C_{u_i} \setminus i} I(\iota_j = \text{Initial})$$

Here, I is an indicator function which is 1 if its argument is true, and 0 otherwise. u_i is the user associated with action i. The global propensities β and second-level concentration parameter α_0 are part of the HDP, and this probability corresponds to the direct sampling scheme in Teh *et al* [23].

The equivalent probability for the case when $\iota_i = \text{Reinforcement}$ depends only on q^i:

$$p(s_i \mid \iota_i = \text{Reinforcement}, \cdot) = q^i_{s_i} / \sum_j q^i_j \tag{4}$$

In the case that $\sum_j q^i_j$ is 0, $\iota_i = \text{Initial}$ deterministically.

This concludes the sampling scheme for each ι_i: evaluate Equation (1) using (2), (3), (4), and Algorithm 2, then draw a Bernoulli random variable according to that probability. This forms the bulk of the inference procedure. However, we have neglected the global latent variables (learning parameters, HDP parameters) to this point.

Teh *et al*[23] include or give reference to sampling schemes for β, α_0, and γ (the latter being related to our inferences through β), which we use. We do not reproduce them here; see Teh *et al* and Escobar and West [4] for details.

This leaves the global learning parameters ϕ and ϵ, and the feature weights of the reward function R. We sample these parameters using Metropolis-Hastings: proposals are generated from a proposal distribution (we use a Gaussian), then accepted or rejected based on the probability of the proposed parameter and the proposal distribution. This allows us to indirectly sample from the full conditional distributions of these parameters. For example, consider the forgetting parameter ϕ:

$$p(\phi \mid s, \iota, \cdot) = \frac{p(s, \iota \mid \phi, \cdot)p(\phi)}{\int p(s, \iota \mid \phi', \cdot)p(\phi')d\phi'} \tag{5}$$

Where ι and s are the vectors of action-specific indicators and action types respectively. $p(s, \iota \mid \phi, \cdot)$ is easy to compute with Algorithm 2 and Equations (2) and (4), but the resulting distribution is difficult to sample from directly. Instead, we generate a proposal ϕ', and accept that proposal ($\phi \leftarrow \phi'$) with probability:

$$\min\left(1, \frac{p(\phi' \mid s, \iota, \cdot)p(\phi' \to \phi)}{p(\phi \mid s, \iota, \cdot)p(\phi \to \phi')}\right)$$

$p(\phi \to \phi')$ is the probability of moving from ϕ to ϕ' using the proposal distribution. The integral in the denominator of Equation (5) cancels, so sampling is as easy as rerunning Algorithm 2 for each user. The remaining learning parameters can be sampled likewise.

The overall inference procedure is then Gibbs sampling: pick initial values for the latent variables, then sequentially sample from each full conditional distribution. Repeating this sequential sampling, the procedure draws from the posterior distribution given our observations in the limit, and in practice is a good approximation (using a finite number of samples) after a burn-in period.

4. EXPERIMENTS

With a model and inference algorithm, we turn to empirical questions: how much data is required to recover the parameters? Is the model useful for describing real data? What can we learn from it?

4.1 Data

We collected a set of 174783 submissions and comments on submissions by 1696 users from the social media website Reddit, along with 2024160 related comments and submissions from other users which we use to compute reply counts. The comments and submissions are across 7037 "subreddits", communities of varying sizes which compose Reddit. These subreddits are the actions in our model: users select between them when choosing to post on a *new* submission. We group together comments by the same user on the same submission, and this grouping is reflected in the count of 174783 comments and submissions. Users were selected through a crawling process, excluding self-identified bots.

distribution over initial and reinforcement distributions, is exactly equivalent to the more standard one in which each contribution comes from the "reinforced" version of the initial distribution.

4.2 Features

One important consideration is feature extraction: what is the exact form of the reinforcement function R? The main social-psychological reward features we use are relative reply counts and relative "karma", the latter being the result of other users voting a comment or submission up or down. These features are first transformed into quantiles (between 0 and 1) of the karma and reply counts respectively across all the data we collected.[3] When grouping contributions on the same submission, we pick the contribution with the most prominent "level", breaking ties by picking the user's first comment on that submission. Submissions are the first level, followed by top-level comments to submissions, replies to those comments, and so on. Huberman *et al* [10] find that more experienced users eventually measure feedback relative to their own previous contributions rather than contributions by other users; this is an interesting direction for more complex future models.

We also include three binary features which indicate the prominence of the contribution. These binary features are 1 respectively if (1) the contribution is a submission, (2) a reply to a submission, or (3) a reply to another comment, and 0 otherwise. These features serve as intercepts for the regression (they are mutually exclusive). To summarize, we fit the following reinforcement utility function:

$$\begin{aligned} R(r) =& A \cdot r_{\text{karma}} + B \cdot r_{\text{replies}} + C \cdot I(r_{\text{type}} = \text{Submission}) \\ &+ D \cdot I(r_{\text{type}} = \text{Top}) + E \cdot I(r_{\text{type}} = \text{Reply}) \end{aligned}$$

With observed feature vector r and regression coefficients A-E.

4.3 Priors and parameters

We perform inference on all of the model parameters rather than fixing their values, but must first specify prior distributions for each. Using shape and rate parameters for Gamma distributions, we have weak priors $\text{Gamma}(1, 0.1)$ and $\text{Gamma}(1, 1)$ on the HDP parameters γ and α_0 respectively as in Teh *et al* [23]. For both learning parameters ϕ and ϵ, we use the prior $\text{Beta}(1, 9)$. The reward feature coefficients and intercepts have priors $\text{Gamma}(1.5, 4)$.

After a burn-in period of 3000 samples, we average predicted distributions across 250 samples, skipping 30 iterations between each to avoid storing and processing correlated samples. For Metropolis-Hastings samples, we use Gaussian proposals with standard deviation $\sigma = 0.01$, except for ϕ and ϵ, for which we use $\sigma = 0.005$ to avoid excessive rejections.

4.4 Scoring probabilistic predictions

What is the right way to score predictions about a user's next subreddit choice? There is only one "correct" choice, corresponding to what the user actually does, but this choice is dependent on many unobserved external factors (perhaps unobservable under some conceptions of free will). Rather than attempting to make a single prediction, we focus on quantifying our uncertainty.

Given a prediction and an event's true outcome, a scoring rule quantifies the performance of that prediction. For one well known class of scoring rules, *strictly proper* scoring rules [21], an agent maximizes her score in expectation by truthfully revealing her beliefs. However, we are interested in comparing the performance of several different models. Strictly proper scoring rules have an associated *divergence function* [7], which measures the divergence of a predicted distribution from an unknown true distribution.

We assume that there is a distribution S over states of the world, in our case encompassing a user's entire history and current state of mind. For a given state of the world $\sigma \sim S$ there is a true distribution over observations f_σ and a predicted distribution g_σ, i.e. a model's probabilistic prediction. Based on a set $X = \{x_1, \ldots, x_N\}$ of observations (i.e. subreddits) drawn from the mixture distribution $\sigma_i \sim S, x_i \sim f_{\sigma_i}$, we want to estimate an expected divergence $E_{\sigma \sim S}[d(f_\sigma || g_\sigma)]$. In other words, how well does the model predict the true distribution of observations? If d is the divergence function of strictly proper scoring rule Q, $\frac{1}{N}\sum_{i=1}^{N} Q_{x_i}(g_{\sigma_i})$ approximates this expectation up to a constant (a generalized entropy term depending only on the true distribution). We use the quadratic scoring rule, which has divergence function $d(f||g) = ||f - g||_2^2$. Comparing models empirically with the quadratic scoring rule compares their predictions' expected squared Euclidean distance from the unknown true distribution of observations.

The quadratic scoring rule is computed as:

$$Q_i(p) = 2g_i - \sum_j g_j^2 \tag{6}$$

Where i is the true outcome, and g a vector of predicted probabilities. The score ranges from 1 when all probability mass is placed on the true outcome, to -1 when all probability mass is placed on an incorrect outcome.

4.5 Models and baselines

Reinforcement is the full model described previously, where users update their propensities in response to social feedback.

UserAll predicts that a user posts in a subreddit proportional to the number of times he or she has done so previously.

UserKMax predicts a subreddit will be chosen next proportional to the number of times it was chosen by the user in that user's past K contributions, maximizing over K. On the real data, this is achieved at approximately $K = 20$; we omit this baseline for synthetic data to simplify presentation.

Global predicts that users pick a subreddit proportional to the number of times it has been picked globally (across all users).

ErevRoth removes the learned initial propensity model, assuming instead that q^0 (Algorithm 1) is uniform over communities.

Initial removes the learning aspects of the reinforcement model (setting the reinforcement function R to 0 deterministically), leaving only the initial propensities. This model smooths a user's local preferences by incorporating global popularity.

InitKMax like UserKMax, trains the initial propensity model on the past K comments from each user, maximizing over K ($K = 25$ on the real data).

True For synthetic data, subreddits are drawn from a true distribution, which serves as an omniscient baseline for that data (but is unfortunately unavailable for real data).

4.6 Performance

Figure 2 shows the performance of the reinforcement model and a variety of baselines on held-out real and synthetic data. The parameters used to generate the synthetic data were chosen to approximately match those inferred from the real data. The performance of the inference algorithm is almost exactly the same as that of the true distribution of held-out subreddit choices on synthetic data. This indicates that the inference algorithm is effective, and that there is enough data available to make accurate inferences (the synthetic and real datasets are approximately the same size).

Turning to real data, we see very similar relative performance. The reinforcement model comes closest to the true distribution of subreddit choices. Simple baselines such as UserAll perform fairly well considering that they are static models, not allowing for behavior changes over time. The UserKMax baseline attempts to com-

[3] We considered using quantiles within a subreddit and contribution level, but analogs of Figure 1 indicate that the data does not support this grouping.

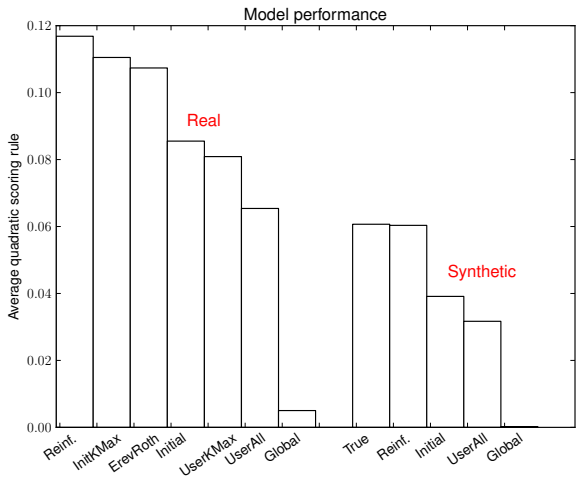

Figure 2: Performance of the models on real and synthetic data. Performance is averaged over the last 7 comments from each user, with each being respectively held out (along with all following comments) and its distribution predicted, among users with at least 8 comments (most having significantly more).

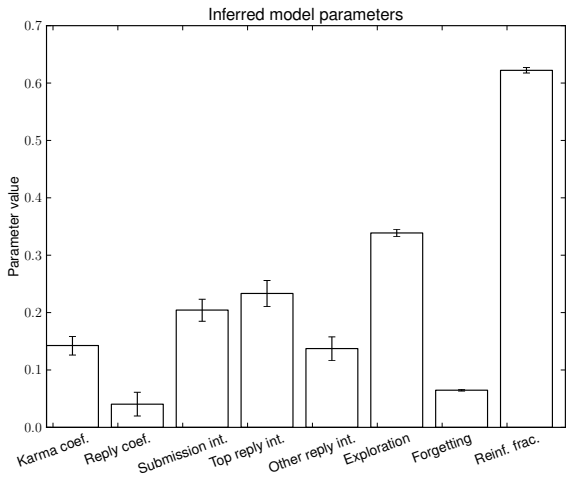

Figure 3: Inferred parameter values using the reinforcement model on real data. Reinforcement function coefficients, intercepts, learning parameters, and the fraction of contributions which were inferred to be the result of reinforcement (ι indicators) are shown. Error bars show empirical 95% credible intervals (contiguous about the mean).

pensate for this by throwing out older data, and does offer a significant improvement over the static UserAll baseline. While predictions based on global subreddit popularity are not very accurate alone, smoothing the user baseline by including global subreddit popularity turns out to be quite effective, as evidenced by the initial propensity baseline. As with the user baseline, we can attempt to adapt the static method to this dynamic setting. InitKMax is the best performer of the non-learning baselines, but the performance loss compared to the reinforcement model is statistically significant ($p = 2.8 \times 10^{-31}$ using a paired t-test).

Figure 4 shows the performance of the reinforcement and initial propensity models as a function of the amount of data they are trained on, on both synthetic and real data. On synthetic data, the reinforcement model quickly approaches the true distribution, while the initial propensity baseline peaks and then declines as agents change their behavior in response to (simulated) feedback. The performance of these two models on real data is strikingly similar to their performance on the synthetic data: the reinforcement model quickly climbs and then stabilizes, while the initial propensity model peaks and then declines as it is "weighed down" by older data. Maintaining a moving window (as in InitKMax) removes the decline, partially compensating for unmodeled user learning.

While sub-optimal, this sliding window in combination with our initial propensity model does quite well despite ignoring a learning dynamic unquestionably present in the synthetic data and, considering Figure 1, also in the real data. However, the sliding window model does not offer a generative model of behavior or any explanatory power; by contrast, the reinforcement model actually explains the process by which changes in behavior take place, rather than just playing catch-up with observed behavior changes. We explore the value of this explanatory power in the next two sections.

4.7 Inferred parameters

Now that we have established that the reinforcement learning model is useful for describing real data, what can it tell us about human behavior? Figure 3 shows the inferred parameters.

Exploration is fairly high, at about 34%. This is sensible for an environment like social media where diversity of interests is critical. Even if rewards are heavily concentrated in a small number of subreddits, users will still pursue a broad range of interests. At the same time, the majority of contributions (about 62%) are modeled as previous experiences rather than initial propensities.

Perhaps the most interesting insights come from the reinforcement function R: what motivates people in social media? One surprise is the relative prominence of the intercept, applied regardless of the social-psychological feedback received in response to a contribution. Interpreting the intercept presents a puzzle: does the contribution simply provide information about a behavior change which has already happened, or does the user's behavior change as a result of having made that contribution? This is an extremely difficult question to answer in general, but we can provide some related insights which hint toward the latter explanation.

We split the intercept in the reinforcement model into three contribution levels: (1) submissions, (2) comments which are top-level replies to submissions, and (3) replies to other comments. Submissions and their replies (levels (1) and (2)) are far more visible: anyone visiting the subreddit will see contributions in level (1), and anyone who looks at the comment thread for a submission will see comments in level (2). Replies to other comments (level (3)) are much less prominent, and are sometimes not displayed on the main comment page at all without additional user interaction. The inferred intercept in case (3) is significantly lower than those inferred for cases (1) and (2). Under a causal interpretation, this might correspond to a higher level of social commitment associated with cases (1) and (2), and therefore a correspondingly more pronounced behavior change. Under the non-causal interpretation, an alternative hypothesis is that users who have already committed more heavily to a specific subreddit are more likely to make prominent contributions to that subreddit.

In either case, social feedback in the form of voting and replies is of greater relative importance when the user's contribution is a reply to a comment than it is for submissions and top-level comments. Surprisingly, the voting score of a contribution is more significant

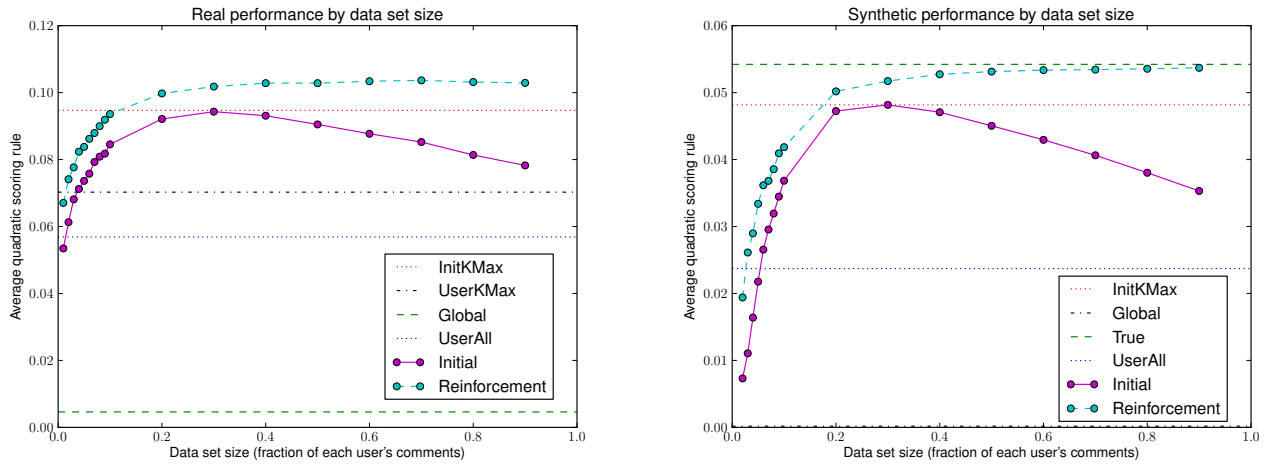

Figure 4: Performance of the reinforcement and initial propensity models as the amount of data varies, for both real and synthetic data. Performance is measured on held-out test data, which is always each user's last contribution. In order to truncate the data, we remove earlier contributions by each user first, leaving a contiguous training set directly before the test data. There are approximately 150000 contributions in the training set for the synthetic data, and about 170000 for the real data.

than the number of replies that contribution receives. One potential explanation is that replies are not always positive, while a high voting score is a clear indicator of community approval.

The reinforcement model enables a form of regression for user behavior changes. Here we have included two social feedback features (voting and reply counts), but other features are possible. Not only can we predict changes in user behavior, but we are also able to articulate specific reasons for those behavior changes.

4.8 Seeding communities

How do you start a social news site from scratch? If no one is participating, new users will be turned off by a lack of activity and content. Reddit's founders faced this problem, and solved it by posting content from fake accounts for the first few weeks of Reddit's existence [17]. Social feedback exhibits a similar issue: without existing users to provide the feedback, new users will not receive enough interaction to keep them interested. Given that users go where the feedback is, how do you start a community from scratch? One answer is providing feedback through an initial set of "seed" participants, who may be sybils or paid participants. We simulate the effects of such seed participants on a group of agents, using the generative model of user behavior learned with the reinforcement model to better understand the dynamics.

Consider four communities A, B, C, and D with a common user base of 100 users. For simplicity, each user has identical initial propensities $(0.3, 0.3, 0.3, 0.1)$, participating in community D with probability 0.1. Users take turns, selecting communities according to our reinforcement model using the same parameters inferred from real data above. Having selected a community, they reply to a random *new* comment (since their last visit). In that same community, they provide positive feedback via voting to each new comment independently with probability 0.3. Each reply or vote increases the associated feedback feature (element of r_i) by 1; since users receive feedback in every community, the magnitude of this feedback quickly becomes irrelevant, only its relative frequency affecting participation probabilities.

Under this model, participation in community D *decays* from its initial proportion of 0.1 to about 0.04: users get more reinforcement in other communities, and so only visit D because a third of the feedback they receive goes to their initial propensities ($\epsilon = 0.33$). This result is quite stable: D has almost no chance of growing. Is it possible to seed the community?

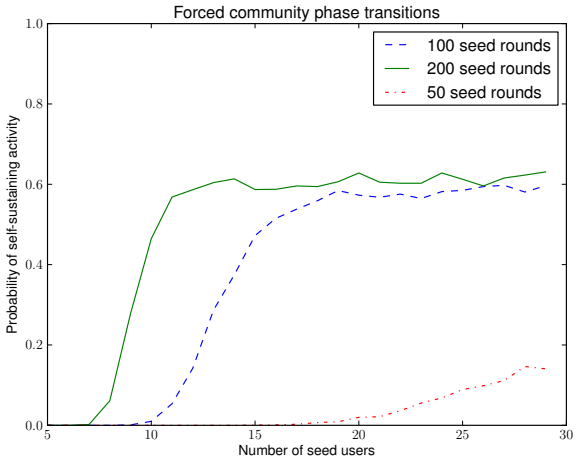

Figure 5: Seeding success probability depends heavily on the number of users doing the seeding, and on the time spent.

We add K seed users who participate only in community D, providing 1 voting feedback to every new comment during their turn. Their purpose is to make D a *self-sustaining* active community by providing extra social feedback during an initial seed period. The game proceeds in rounds, with every user taking one turn during each round in a consistent order. The seed users participate for the first S rounds. We consider D to be self-sustaining and active if the average non-seed user spends 50% of their time in community D after an additional 500 rounds without any seed users.

Figure 5 shows the probability of successfully seeding community D as a function of the number of seed users K and the number

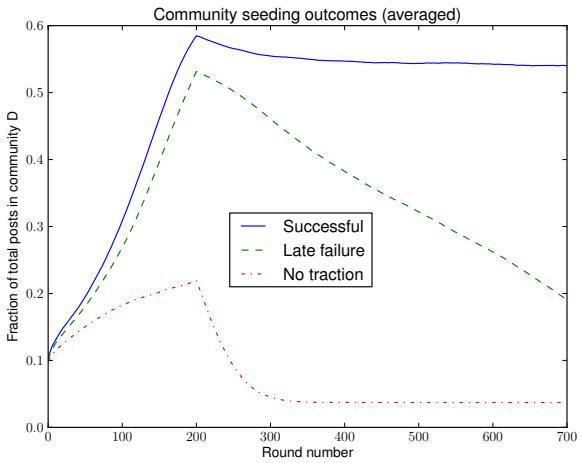

Figure 6: Three types of observed outcomes in synthetic community seeding experiments, with 200 seed rounds and 9 seed users. Sequences were grouped based first on the fraction of interest in community D at round 200: no traction (≤ 0.4) or some early traction. Of those with early traction, there are late failures (≤ 0.5 at 700) and successfully seeded communities. Curves are averaged within each group.

of seed rounds S. Even a large number of users has a small chance of seeding a community in a short time, but relatively small numbers of users over a long period of time can force phase transitions. Figure 6 shows example dynamics of three common outcomes of the seeding process: no traction, late failure, and successful seeding of a self-sustaining community.

5. CONCLUSIONS

We have shown that a simple model of learning can capture complex behavior changes in social media. Users spend more time in communities where they have received social-psychological feedback, and in communities where they have previously invested more time. While behavior is stochastic, an analogy to humans playing mixed strategies in matrix games provides a simple and effective learning model in this setting. Our quantitative model gives insight into individual user behavior in social media, and provides a solid foundation for studying the dynamics of communities of agents with mutual feedback and complex collective learning.

6. ACKNOWLEDGMENTS

This work was supported in part by a US National Science Foundation (NSF) CAREER award (IIS-1414452), and in part by NSF grant IIS-1124827.

7. REFERENCES

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. ICML*, page 1, 2004.

[2] S. Chernova and M. Veloso. Confidence-based policy learning from demonstration using Gaussian mixture models. In *Proc. AAMAS*, pages 233:1–233:8, 2007.

[3] I. Erev and A. E. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Amer. Econ. Rev.*, 88(4):848–81, September 1998.

[4] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *J. ASA*, 90(430):577–588, 1995.

[5] K. Genter, N. Agmon, and P. Stone. Ad hoc teamwork for leading a flock. In *Proc. AAMAS*, pages 531–538, 2013.

[6] E. Gilbert. Widespread underprovision on Reddit. In *Proc. CSCW*, pages 803–808, 2013.

[7] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *J. ASA*, 102(477):359–378, 2007.

[8] C. Heaukulani and Z. Ghahramani. Dynamic probabilistic models for latent feature propagation in social networks. In *Proc. ICML*, pages 275–283, 2013.

[9] G. Hsieh, Y. Hou, I. Chen, and K. N. Truong. "Welcome!": Social and psychological predictors of volunteer socializers in online communities. In *Proc. CSCW*, pages 827–838, 2013.

[10] B. A. Huberman, D. M. Romero, and F. Wu. Crowdsourcing, attention and productivity. *J. Inf. Sci.*, 35(6):758–765, 2009.

[11] C. L. Isbell, M. Kearns, S. Singh, C. Shelton, P. Stone, and D. Kormann. Cobot in LambdaMOO: An adaptive social statistics agent. *JAAMAS*, 13(3):327–354, 2006.

[12] W. B. Knox and P. Stone. Combining manual feedback with subsequent MDP reward signals for reinforcement learning. In *Proc. AAMAS*, pages 5–12, 2010.

[13] W. B. Knox and P. Stone. Reinforcement learning from simultaneous human and MDP reward. In *Proc. AAMAS*, pages 475–482, 2012.

[14] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proc. WWW*, pages 621–630, 2010.

[15] K. Lerman and T. Hogg. Using stochastic models to describe and predict social dynamics of web users. *ACM Trans. Intell. Syst. Technol.*, 3(4):62:1–62:33, Sept. 2012.

[16] K. Lerman, S. Intagorn, J.-H. Kang, and R. Ghosh. Using proximity to predict activity in social networks. In *Proc. WWW*, pages 555–556, 2012.

[17] K. Morris. How Reddit's cofounders built Reddit with an army of fake accounts. *The Daily Dot*, June 2012.

[18] L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[19] M. Munie and Y. Shoham. Joint process games: From ratings to wikis. In *Proc. AAMAS*, pages 847–854, 2010.

[20] A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In *Proc. ICML*, pages 663–670, 2000.

[21] L. J. Savage. Elicitation of personal probabilities and expectations. *J. ASA*, 66(336):783–801, 1971.

[22] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Comm. ACM*, 53(8):80–88, 2010.

[23] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. ASA*, 101(476), 2006.

[24] F. Wu and B. A. Huberman. Novelty and collective attention. *PNAS*, 104(45):17599–17601, 2007.

[25] F. Wu, D. Wilkinson, and B. Huberman. Feedback loops of attention in peer production. In *Proc. CSE*, volume 4, pages 409–415, 2009.

[26] H. Zhu, A. Zhang, J. He, R. E. Kraut, and A. Kittur. Effects of peer feedback on contribution: A field experiment in Wikipedia. In *Proc. CHI*, pages 2253–2262, 2013.