A Generative Model for Discovering Action-Based Roles and Community Role Compositions on Community Question Answering Platforms

Chase Geigle, Himel Dev, Hari Sundaram, ChengXiang Zhai
University of Illinois at Urbana-Champaign
geigle1@illinois.edu, dev3@illinois.edu, hs1@illinois.edu, czhai@illinois.edu

Abstract

This paper proposes a generative model for discovering user roles and community role compositions in Community Question Answering (CQA) platforms. While past research shows that participants play different roles in online communities, automatically discovering these roles and providing a summary of user behavior that is readily interpretable remains an important challenge. Furthermore, there has been relatively little insight into the distribution of these roles between communities. Does a community's composition over user roles vary as a function of topic? How does it relate to the health of the underlying community? Does role composition evolve over time? The generative model proposed in this paper, the mixture of Dirichlet-multinomial mixtures (MDMM) behavior model can (1) automatically discover interpetable user roles (as probability distributions over atomic actions) directly from log data, and (2) uncover community-level role compositions to facilitate such cross-community studies.

A comprehensive experiment on all 161 non-meta communities on the StackExchange CQA platform demonstrates that our model can be useful for a wide variety of behavioral studies, and we highlight three empirical insights. First, we show interesting distinctions in question-asking behavior on StackExchange (where two distinct types of askers can be identified) and answering behavior (where two distinct roles surrounding answers emerge). Second, we find statistically significant differences in behavior compositions across topical groups of communities on StackExchange, and that those groups that have statistically significant differences in health metrics also have statistically significant differences in behavior compositions, suggesting a relationship between behavior composition and health. Finally, we show that the MDMM behavior model can be used to demonstrate similar but distinct evolutionary patterns between topical groups.

1 Introduction

Discovering user roles and community role compositions on Community Question Answering (CQA) platforms is an important challenge. CQA platforms such as the StackExchange platform[1] play an incredibly important role in today's society, and recent years have seen an increase in both the number of such CQA communities and the user populations within each community. For example, in 2017, StackOverflow[2] added over 200,000 new questions and over 130,000 new users every month; many software developers regularly depend StackOverflow to be effective at work. An understanding of behavior within communities can help to inform the decisions made by platform providers to steer the communities to be maximally effective.

It is well established that users in these communities play important, distinct roles (Adamic et al. 2008; Mamykina et al. 2011; Nam, Ackerman, and Adamic 2009; Wang, Lo, and Jiang 2013; Wu, Baggio, and Janssen 2016), but it remains an important scalability challenge to automatically uncover these distinct user roles across a large number of communities. StackExchange as a platform, for example, facilitates 161 distinct websites. Manual investigation of user behavior compositions within and across these communities is prohibitively expensive to do without some level of automation, and with these communities continuing to grow over time, the need for automated role discovery intensifies.

Existing approaches fall short of our needs in a number of ways. Many existing models for role discovery do not consider the case of modeling many communities at once, yet such a cross-community understanding of behavior is important to enable comparative studies across communities. Previous work often defines roles based on a graph-centric approach (Fisher, Smith, and Welser 2006; Welser et al. 2007; Barash et al. 2009), which fails to uncover many distinct roles beyond "answer people" and "discussion people." Other approaches require a manual definition of individual features to describe roles (Chan, Hayes, and Daly 2010; Furtado et al. 2013), which can fail to cover all of the empirically present role patterns in the data.

In this paper, we propose a generative model for discovering action-based user roles and community role compositions in CQA platforms directly from log data. We formally define an action-based user behavior role as a probability distribution over atomic actions a user may take with respect to the CQA community within one browsing session. We also directly model the role compositions across all communities within the platform to facilitate comparative analysis of communities. This is achieved via the use

[1] https://stackexchange.com

[2] https://stackoverflow.com, the largest community on the StackExchange platform

of a mixture of Dirichlet-multinomial Mixtures (MDMM), which allows us to use statistical inference to uncover the latent user roles and community role compositions from log data directly, which can facilitate studies into user behavior both within and across communities on a CQA platform at scale. We envision that with the assistance of our model, human analysts can "see" more patterns than what they could see otherwise. Such a tool provides a useful "lens" through which to view behavior data, and opens up many directions for future studies that would not otherwise be possible.

To demonstrate that such a model is indeed useful as a tool to assist human discovery of behavior patterns within and between CQA communities, we perform a comprehensive experiment on all 161 non-meta communities on the StackExchange CQA platform that delivers three empirical insights. First, we show interesting distinctions in question-asking behavior on StackExchange (where two distinct types of askers can be identified) and answering behavior (where two distinct roles surrounding answers emerge). Second, we find statistically significant differences in behavior compositions across topical groups of communities on StackExchange, and that those groups that have statistically significant differences in health metrics also have statistically significant differences in behavior compositions, suggesting a relationship between behavior composition and health. Finally, we show that the MDMM behavior model can be used to demonstrate similar but distinct evolutionary patterns between topical groups.

2 Related Work

The presence of roles on CQA platforms has been argued by many. For example, Adamic et al. (2008) demonstrate that, on the Yahoo! Answers platform, there are at least three distinct user types—answerers, askers, and *discussion persons*. Mamykina et al. (2011) argue for the presence of at least four distinct user roles on StackOverflow: *community activists*, *shooting stars*, *low-profile users*, and *lurkers and visitors*. Other studies have explored whether roles characterized by a single action are separate or overlapping (Nam, Ackerman, and Adamic 2009; Wang, Lo, and Jiang 2013). Developing tools to automatically uncover distinct user behavior types is a major thrust of this paper.

Many approaches for discovering these distinct user roles in the CQA setting require practitioners to define individual features used to describe the discovered roles (Chan, Hayes, and Daly 2010; Furtado et al. 2013), and early work in the domain of user role modeling could only easily identify two critical roles ("answer people" and "discussion people") through the use of a graph-centric modeling approach (Fisher, Smith, and Welser 2006; Welser et al. 2007; Barash et al. 2009). More recent work explores a mixed-membership approach to user behavior modeling (White et al. 2012) in order to identify more user roles, but still takes a graph-centric modeling approach. In this work, we explore the newer direction of action-focused probabilistic modeling for user behavior in order to automatically discover roles in a way that requires less hands-on effort to define features and is flexible enough to be able to capture more nuance within the roles of "answer people" and "discussion people".

The application of probabilistic modeling for user behavior understanding has been explored before (Manavoglu, Pavlov, and Giles 2003; Xu et al. 2012; Qiu, Zhu, and Jiang 2013). We extend this body of research by modeling the behavior composition at a community level, rather than just at a user level. This allows us to understand the behavior at the level of an entire community as it relates to others.

Perhaps the most relevant probabilistic behavior model to ours is the one proposed by Han and Tang (2015), where they attempt to jointly model three phenomena: social network link formation, community discovery, and behavior prediction. Their definition of user behavior differs from ours, however, as it considers only posting and reposting as the two possible actions a user can take. We attempt to define a much more comprehensive behavioral action set in this work. Furthermore, their discovered role distributions model real-valued user attributes, rather than behavior directly, which makes interpretation challenging. Our work, in comparison, assumes a different generative process over user action lists that leads to a set of readily interpretable probability distributions that define our roles.

CQA data, and in particular the StackExchange CQA platform, have been analyzed in many ways in previous literature (Nam, Ackerman, and Adamic 2009; Mamykina et al. 2011; Adamic et al. 2008; Wang, Lo, and Jiang 2013; Furtado et al. 2013; Anderson et al. 2012), but many do not discuss user roles in depth. Furtado et al. (2013), however, do explore user roles and their dynamics using five of the communities on the StackExchange platform, but their definition of user roles arises from manual construction of user attributes and an agglomerative clustering approach. Our model, in comparison, is more general in that it should be applicable to any CQA community (or any social network) where articulating the set of actions users can take within the community is the only manual supervision required.

Our session-focused approach is closely related to the notion of clickstream mining (Wang et al. 2016; Gündüz and Özsu 2003; Geigle and Zhai 2017; Su and Chen 2015; Benevenuto et al. 2009; Lu, Dunham, and Meng 2006; Sadagopan and Li 2008), where a variety of clustering techniques is applied to find users that share similar clickstream traces. Many of these techniques utilize Markov models and focus on the task of predicting a user's next action. In this paper, we instead focus on characterizing the behavior of users in an interpretable way that also facilitates cross-community comparisons.

The model we propose in this paper is essentially similar to topic models such as PLSA (Hofmann 1999) or LDA (Blei, Ng, and Jordan 2003), but the key difference is that the data modeled by our model are the user actions whereas topic models generally model text data where the input tokens are individual words within topics. The Dirichlet-multinomial mixture (DMM) (Nigam et al. 2000; Yin and Wang 2014) is the closest related model to ours in this space. A DMM assumes that individual documents exhibit only one topic—our generative framework also assumes that one user session exhibits only one role.

Other approaches for user behavior modeling on CQA communities consider both actions and textual content

to generate topic-specific action distributions (Qiu, Zhu, and Jiang 2013; McCallum, Wang, and Corrada-Emmanuel 2007). These distributions are similar to what we call roles, but the meaning they capture is very different—in their work these capture how users interact with a specific topic, whereas in our work they describe how to characterize an individual user's entire browsing session.

3 Model

The design of our model is motivated by our goal of discovering interpretable descriptions of functional roles played by users on CQA platforms, as well as a representation for each community as a mixture over these user roles. We explore a definition of user roles that considers the co-occurrence behavior of actions users take within individual browsing sessions. To accomplish this, we represent the roles as probability distributions that describe the likelihood of taking individual actions when a user is assuming a particular functional role in one session. This definition is advantageous: first, it is general, and thus should be applicable to any CQA platform (or even any social network); second, roles represented in this way can be readily interpreted by inspection; and third, it is able to capture the uncertainty associated with assigning users to roles.

3.1 Generative Process and Inference

The first step in the use of our action-based role discovery model is to define the set A of actions users may take within a community. Defining the actions in this action set is very important in order to capture meaningful roles under our model, so careful attention should be paid to the construction of a set of disjoint actions whose proportions can meaningfully reflect a type of domain-relevant behavior.

Next, one must identify the collection of *observed* communities $C_{1:N}$ to analyze that all share the same action set A. We do not address the problem of community discovery in this paper; rather, these communities are treated as input to the model. Each community must share the same types of allowed actions. In our case, we use individual websites that are all part of the same CQA platform (but focus on different topical domains) to ensure that by defining A with respect to the CQA platform itself we can represent behavior across all of these communities.

To automatically discover distinctive user behavior types, which we will call our roles, we appeal to the general technique of probabilistic graphical models (Koller and Friedman 2009) and model user behavior using a mixed membership approach. The model assumes that there are K distinct user roles, each of which is characterized with a categorical distribution ϕ_k over actions from some A; each of the roles ϕ_k is assumed to be drawn from a Dirichlet distribution with parameter β. With these user roles defined, we further assume that each community C_i is associated with a mixing distribution θ_i (drawn from another Dirichlet distribution with parameter α) that governs the distribution over the user roles for each *user session* that occurs *within that community*. If a user makes actions in multiple communities within one browsing session, we subdivide their browsing session into a collection of sessions, one for each community they participated in.

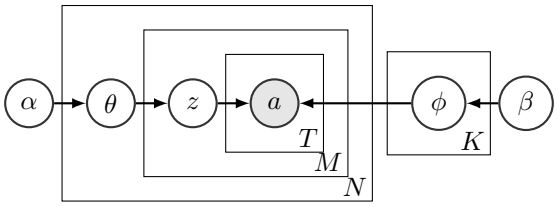


Figure 1: Plate notation for the role discovery model we propose. α parameterizes a Dirichlet distribution from which each community's role proportions, θ_i, are drawn. z represents the role assignment for a specific user session, and a represents the actions taken within that user session. β parameterizes a Dirichlet distribution from which each of the user roles ϕ_k are drawn, each of which is a categorical distribution over the possible action types.

More concretely, we represent each community C_i with a list of the user sessions $\langle \mathbf{s_{i,1}}, \mathbf{s_{i,2}}, \ldots, \mathbf{s_{i,M}} \rangle$ associated with it. Each session is itself a list of actions $\mathbf{s_{i,j}} = \langle a_{i,j,1}, a_{i,j,2}, \ldots, a_{i,j,T} \rangle$, with each $a_{i,j,t} \in A$. Each individual session $\mathbf{s_{i,j}}$ is associated with one particular user role $z_{i,j}$ that indicates the role distribution $\phi_{z_{i,j}}$ from which each of the actions within the session is drawn (note that an individual user is free to exhibit a different roles between different sessions). The full generative process is thus

1. For $k = 1$ to K (number of roles), draw an action distribution $\phi_k \sim \text{Dirichlet}(\beta)$
2. For each community C_i:
   (a) Draw a role mixing distribution $\theta_i \sim \text{Dirichlet}(\alpha)$
   (b) For each $\mathbf{s_{i,j}}$ in community C_i:
      i. Draw a role for the session $z_{i,j} \sim \text{Categorical}(\theta_i)$
      ii. For $t = 1$ to $|\mathbf{s_{i,j}}|$ (the length of the session), draw a single action within the session $a_{i,j,t} \sim \text{Categorical}(\phi_{z_{i,j}})$

and is depicted using plate notation in Figure 1.

The resulting model is quite similar to a Dirichlet-multinomial mixture (DMM), which has seen use in the text mining community for clustering (Yin and Wang 2014) and classification (Nigam et al. 2000). A major difference from our model, however, is that in a DMM one learns a *single* distribution θ that governs the mixing proportions over the components ϕ_k that is shared for each element C_i, whereas our model instead learns a *separate* distribution θ_i for each individual community, but shares the description of the components ϕ_k between each. This allows us to compare two communities by their role proportions in a meaningful way since each θ_i will be a distribution over the same set of roles ϕ_k. If one were instead to fit multiple DMMs, one for each community, comparison of the θ distributions would not necessarily be immediately obvious due to the fact that each model would learn a separate set of roles ϕ_k. Thus, we view our model as a principled mixture of DMMs (MDMM) where we have made a deliberate decision to share a global set of role components between all communities C_i.

There are several approaches to inference in a DMM. Nigam et al. (2000) use maximum a posteriori (MAP) estimation to obtain a point estimate. We instead choose to follow a more fully Bayesian approach similar to Yin and Wang (2014) and instead appeal to Markov-chain Monte Carlo methods to approximate the desired posterior distribution. Specifically, we integrate out θ and ϕ in order to then derive a collapsed Gibbs sampler that iteratively updates the latent role assignments z_j by sampling new values from the full conditional distribution. When this chain has converged, we extract a MAP estimate for each θ_i and ϕ_k from the current state of the Markov chain.

Formally, we can define the full conditional distribution

$$p(z_{m,n} = z \mid \mathbf{Z}_{\neg m,n}, \mathbf{S}, \alpha, \beta) = \frac{p(\mathbf{Z}, \mathbf{S} \mid \alpha, \beta)}{p(\mathbf{Z}_{\neg m,n}, \mathbf{S} \mid \alpha, \beta)} \propto \frac{p(\mathbf{Z}, \mathbf{S} \mid \alpha, \beta)}{p(\mathbf{Z}_{\neg m,n}, \mathbf{S}_{\neg m,n} \mid \alpha, \beta)}, \quad (1)$$

where $\mathbf{Z}_{\neg m,n}$ indicates the set of all the assignments of $z_{i,j}$ with only $z_{m,n}$ excluded, and similarly $\mathbf{S}_{\neg m,n}$ indicate the set of all user sessions with only the specific session $\mathbf{s_{n,m}}$ absent. We begin by noting $p(\mathbf{Z}, \mathbf{S} \mid \alpha, \beta) = p(\mathbf{S} \mid \mathbf{Z}, \beta)P(\mathbf{Z} \mid \alpha)$, and focus on each term separately. Following a similar argument to Yin and Wang (2014), we have $p(\mathbf{Z} \mid \alpha) = \prod_{i=1}^{N} \frac{B(\eta_i + \alpha)}{B(\alpha)}$, where $B(\alpha)$ is the multivariate beta function and η_i is a vector where $\eta_{i,k}$ indicates the number of times role k is chosen as the role assignment for a session in community C_i. Similarly, $p(\mathbf{S} \mid \mathbf{Z}, \beta) = \prod_{k=1}^{K} \frac{B(\tau_k + \beta)}{B(\beta)}$ where τ_k is a vector with $\tau_{k,a}$ indicating the number of times action type a was assigned to role k through its session's role assignment. From here, we can derive the sampling probability through cancellation of terms and exploiting the property of the gamma function that $\Gamma(1 + x) = x\Gamma(x)$ and arrive at

$$p(z_{m,n} = z \mid \mathbf{Z}_{\neg m,n}, \mathbf{S}, \alpha, \beta) \propto \frac{\alpha_z + \eta_{i,z}^{\neg m,n}}{\sum_{k=1}^{K} \alpha_k + \eta_{i,k}^{\neg m,n}} \times \frac{\prod_{a \in \mathbf{s_{m,n}}} \prod_{j=1}^{c(a, \mathbf{s_{m,n}})} (\beta_a + \tau_{z,a}^{\neg m,n} + j - 1)}{\prod_{j=1}^{|\mathbf{s_{m,n}}|} \left(\left(\sum_{a=1}^{|A|} \beta_a + \tau_{z,a}^{\neg m,n} \right) + j - 1 \right)}, \quad (2)$$

where $c(a, \mathbf{s_{m,n}})$ indicates the number of occurrences of action type a in session $\mathbf{s_{m,n}}$.

As a practical matter, computing this probability is susceptible to underflow issues due to the products occurring in the second term. To prevent this issue, we use the Gumbel-max trick (Maddison, Tarlow, and Minka 2014) to sample from this discrete distribution. This trick works by first computing the sampling proportions in log-space $\gamma_k = \log \tilde{p}(z_{m,n} = k \mid \mathbf{Z}_{\neg m,n}, \mathbf{S}, \alpha, \beta)$, where $\tilde{p}$ represents the un-normalized probability in equation 2, which effectively prevents the underflow issues. We then can sample from the original discrete distribution by sampling k values $g_k \sim \text{Gumbel}(0)$, and taking the sample $z_{m,n} = \text{argmax}_k \, \gamma_k + g_k$. We have open-sourced the implementation of our inference algorithm under a liberal license[3].

[3]https://github.com/CrowdDynamicsLab/stackoverflow-stream

3.2 Choosing the Number of Roles

The number of roles, K, remains a hyperparameter of the MDMM behavior model. How should one choose the "optimal" value for K? This is a similar question that is asked for nearly any mixed-membership or clustering model. We note, first, that the choice of K can be an empirical parameter that is sometimes beneficial as it can give users *control* over the granularity of the model, much like a user can adjust the zoom level of a microscope. If the user does not know how to set K a priori, we describe a procedure that can help choose a particular value of K that may be optimal.

In our specific case, not only do we wish to discover distributions over actions that can adequately describe a user's behavior within a single session, but we wish for these distributions to be *meaningfully different* from one another. An ad-hoc approach, then, is to simply run the model for different values of K in some range, and then investigate the roles $\phi_{1:K}$ that are produced. When moving from k to $k+1$ roles, if a new role arises that is not meaningfully different from all of the k roles found previously, this suggests that k was the optimal number of roles for the data being modeled.

One can define a simple quantitative heuristic to capture this intuition. Formally, let $\phi_{1:k}$ be the k roles proposed by the model previously, and let $\hat{\phi}_{1:k+1}$ be the $k+1$ roles proposed by the model when incrementing K. Consider a single new role $\hat{\phi}_i$. We can compute how different it is from each of the previously proposed roles $\phi_{1:k}$ by using the KL-divergence metric (Kullback 1959). By taking the minimum divergence from the newly proposed role $\hat{\phi}_i$ to each of the k previous roles, we have a measure for how "surprising" this new role is compared to the previous roles. If it is very similar to one of the existing roles, it will have a very low minimum KL-divergence; on the other hand, should it be very different from all of the previous roles, it would have a very large minimum KL-divergence.

If we then take the maximum value of this measure over all of the $k+1$ newly proposed roles $\hat{\phi}_{1:k+1}$, we obtain a number that reflects the largest minimum divergence between the set of k old roles and the set of $k+1$ new roles. The smaller this value, the more redundant the set of $k+1$ new roles is compared to the set of k previous roles. Formally, we can define this measure $\text{MaxMinKL}_{k \to k+1}$

$$\text{MaxMinKL}_{k \to k+1} = \max_{\hat{\phi}_i} \left(\min_{\phi_j} KL(\phi_j \mid\mid \hat{\phi}_i) \right). \quad (3)$$

To find the optimal value of K, one can run the model for K in a range of values to be considered, computing $\text{MaxMinKL}_{k \to k+1}$ for each transition. When this value drops substantially, this is a sign that the new set of roles is not meaningfully different from the previous set of roles, and we should stop increasing K.

3.3 Applications of the Model

The MDMM behavior model is a tool to enable humans to discover new knowledge, explore new hypotheses, and test those hypotheses about user behavior in ways that they were unable to before. There are a number of different applications of the model beyond just the discovery of user behav-

ior roles. We outline a few of them below, but note that this list is not exhaustive—exploring those opportunities are interesting future directions.

Community Profiling. A secondary output of the model are the mixing proportions θ_i over the roles for each community. These distributions provide a profile of the behavior of users within the community, which can be used as a representation for that community in downstream tasks. To explore this in more detail, in Section 4.3 and 4.4 we explore how we can use this representation to uncover communities with different behavior profiles, and show how these groups are correlated with many metrics of community success.

User Profiling. The model can also be used to infer the roles of a user by averaging over the roles they assume in their sessions. This output can then be used in downstream tasks that relate to understanding user behavior on an individual level and can be used as a representation of a user for other machine learning algorithms.

Behavior Dynamics of Communities. We can also uncover temporal community representations by further segmenting the user browsing sessions into buckets relating to different points in time. This allows us to study how behavior proportions evolve over time as community age. We explore this in more depth in Section 4.5.

Behavior Dynamics of Users. In much the same way we can uncover community representations over time, we can also uncover user representations over time. This output could be used to understand how individual users, or groups of users, change their behavior over time.

4 Experiments

The goal of our experiments is to demonstrate the usefulness of the MDMM user behavior model as a tool for investigating user behavior in different ways. Our goal is *not* to be completely comprehensive or conclusive in our study of user behavior, but rather to lay a framework for future studies in a variety of different directions that could not otherwise be studied.

Our MDMM user behavior model provides two important outputs to characterize user behavior in CQA communities: (1) the latent role representations, and (2) the degree to which each latent role is present within each of the CQA communities. We apply our model to communities from the StackExchange CQA platform[4] in order to better understand its utility for role discovery and CQA community behavior analysis tasks. We take the entire StackExchange dataset consisting of a total of 322 websites and discard all "meta" websites (websites discussing one of the other StackExchange websites), leaving us with 161 non-meta websites (communities) for our analysis.

4.1 Dataset Construction

A critical component of the use of the MDMM in our setting is properly defining the action space to be considered, as the roles discovered are to be distributions over that action space. The flexibility of defining actions outside of the MDMM model makes it easy to accommodate analysis of action patterns at different levels of granularity by adjusting the granularity of the action space to be analyzed itself. However, in any specific application, carefully choosing the exact action set used is naturally very important. If the space of actions is defined too narrowly, this prevents discovering subtle differences between user roles.

Table 1: Action names and their definitions for our application of the MDMM behavior model on StackExcahnge. (m: "my", o: "other", q: "question", a: "answer")

Action Name	Action Definition
question	Posting a new question
answer-mq	Answering your own question
answer-oq	Answering someone else's question
comment-mq	Commenting on your own question
comment-oq	Commenting on someone else's question
comment-ma-mq	Commenting on your own answer to your own question
comment-ma-oq	Commenting on your own answer to someone else's question
comment-oa-mq	Commenting on someone else's answer to your own question
comment-oa-oq	Commenting on someone else's answer to someone else's question
edit-mq	Editing your own question
edit-oq	Editing someone else's question
edit-ma	Editing your own answer
edit-oa	Editing someone else's answer
mod-vote	Voting for moderation action
mod-action	Moderating a post

To analyze the StackExchange dataset, we defined an action space based on the inherent content hierarchy present on the StackExchange platform (see Table 1 for a list of the action set we consider). Content on the StackExchange platform comes in three main types: questions (the root content), answers (which nest below questions), and comments (which can nest either beneath questions or answers), so it is natural to consider an action set consisting of the creation action for each of these three types of content. However, limiting the action space to just these three actions will fail to uncover meaningful differences in commenting behavior, the most frequently generated type of content. We subdivide the commenting action by first distinguishing between comments that occur on questions from comments that occur on answers, and then further dividing these based on the original poster of the parent content further up in the content tree. Concretely, we arrive at six separate commenting action types: commenting on my own question (comment-mq), commenting on others' questions (comment-oq), commenting on my answer to my question (comment-ma-mq), commenting on my answer to

[4]The dataset is available here: https://archive.org/details/stackexchange. We used a dataset from 2016-12-12, which covers from 2008-07-31 through 2016-12-11.

others' questions (comment-ma-oq), commenting on others' answers to my question (comment-oa-mq), and finally commenting on others' answers to others' questions (comment-oa-oq). Similarly, we can subdivide the answering action into answering my own question (answer-mq) and answering others' questions (answer-oq).

While creation actions are arguably the most important actions to consider for modeling user behavior with respect to the generation of content, it is also important to consider the role that editors play within the communities. We define four types of edit actions: editing my question (edit-mq), editing others' questions (edit-oq), editing my answer (edit-ma), and editing others' answers (edit-oa). We also include two actions related to moderation (the closing, locking, deleting, moving, etc. of posts) on StackExchange with two actions: voting for moderation activity (mod-vote) and the actual application of moderation (mod-action).

Once we have defined our action space, we can then begin the session segmentation process. We start with a chronologically ordered list of all of the actions from the action space taken within a community, and then partition this list into separate action lists associated with each individual user. Then, we define a session as a contiguous chunk of a user's action list such that the gap between consecutive actions is less than six hours to roughly capture a day's worth of activity per session. The collection of all of these sessions, grouped by community, serves as the MDMM's input.

We further decompose the community session lists by segmenting them into month-long chunks to enable temporal analysis of the behavior compositions over time for our communities. We define the "birth" of a community as the timestamp of the very first action taken in any user session associated with it, and then use that as the reference point for constructing the monthly session lists. This gives us 49,768,660 user sessions across 9117 community-month pairs.

4.2 Analysis of the Discovered Roles

We start our analysis by examining the usefulness of the discovered roles $\phi_{1:K}$. Because the number of roles, K, is a hyperparameter of our model, it must be chosen in advance of our investigation into the roles. Our MaxMinKL heuristic suggests a value of $K = 5$ for our dataset (see Table 2 for the scores for each transition), and manual inspection also indicated role redundancies found at $K > 5$. We ran our model on an Intel(R) Core(TM) i7-5820K CPU, and each iteration takes approximately 20 seconds. We ran the model for 100 total iterations, as we found the output stopped changing appreciably after about 40 iterations. Each role we discovered at $K = 5$ is depicted in Figure 2, along with labels constructed from our own interpretation of the roles. These results directly help us understand what the "typical" roles assumed by users are in CQA communities.

"Eager asker" (Figure 2a): Users exhibiting this role tend to ask questions, and comment on others' answers to their questions.

"Careful asker" (Figure 2d): While both this role and the previous role tend to ask questions in the same proportion within a session, a "careful asker" tends to comment a lot in discussions on their own question rather than on answers to their question, and they also have a much higher chance of updating their question when compared to the "eager asker" role. This subtle difference in asking behavior types would be lost if we had not carefully subdivided the commenting action by considering both the type and originator of the parent content of the comment.

"Answerer" (Figure 2c): For the most part, this reflects a user that is concerned about their own answers. They provide their answers, they comment on their answers, and they update their answers. They may also seek clarification on a question by engaging in the discussion on that question, but not nearly as much as the next role.

"Clarifier" (Figure 2e): Users exhibiting this role tend to engage in the discussion on a question (by far their most frequent action) before answering; they also tend to comment on others' answers to others' questions more than any other role.

"Editor/moderator" (Figure 2b): This role captures nearly all of the observed moderation activity, and the most common action is to update someone else's question.

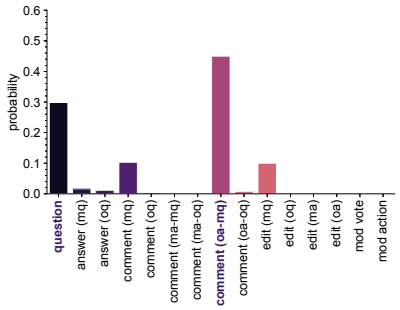
Table 2: The MaxMinKL heuristic for the MDMM behavior model applied to the StackExchange dataset. Notice the substantial drop when moving from $K = 5$ to $K = 6$, indicating redundancy obtained in the set of new roles. This matches our own visual inspection of the role distributions; hence, we choose $K = 5$ for the remaining experiments.

Transition	MaxMinKL
$2 \rightarrow 3$	2.95
$3 \rightarrow 4$	3.35
$4 \rightarrow 5$	3.30
$5 \rightarrow 6$	1.73
$6 \rightarrow 7$	1.75

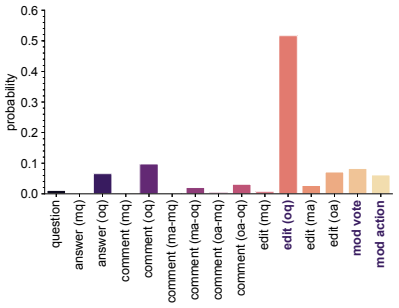
While it might not be very surprising to see two distinct roles corresponding to primarily asking questions and primarily answering questions, the model goes beyond discovering such "obvious" roles to provide further fine distinction of interesting variations of roles for both question askers and question answerers, which may not be easy to discover otherwise by simply manually examining their behaviors. Our MDMM behavior model is able to uncover these meaningful user behavior roles, including those with subtle differences, in a completely unsupervised way directly from log data once given an appropriate action space. Note that due to the generality of the MDMM model, we can easily refine action categories to potentially discover even finer-grained variations of user roles than what we have seen here—in this way, our model naturally supports multi-resolution analysis of user behavior.

4.3 Analysis of Behavior Compositions

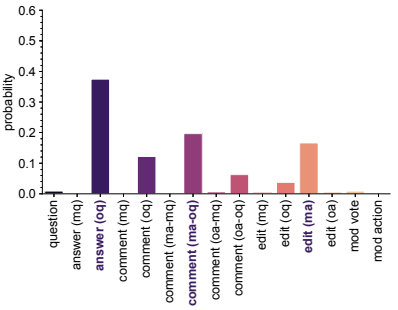
The MDMM behavior model also outputs role proportions θ_i for each community in the dataset. These proportions provide an informative summary of the composition of behav-



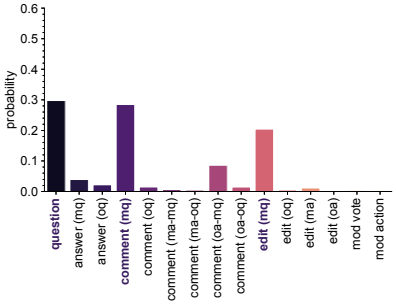
(a) An asker role we call "eager asker." In comparison to Figure 2d, we see that when a user exhibiting this role chooses to comment, they tend to comment on others' answers to their own question.



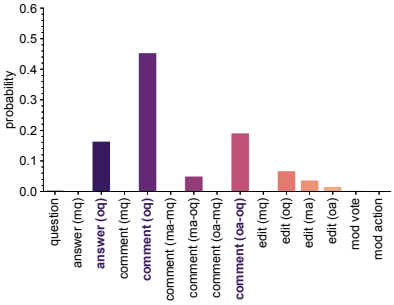
(b) An "editor/moderator" role. This role is the only role that exhibits moderation behavior, and we see that the vast majority of actions a user takes when exhibiting this role are to update others' questions.



(c) An "answerer" role. The majority of the time, users exhibiting this role answer others' questions, engage in discussion on their provided answers, and update their answers accordingly.



(d) An asker role we call "careful asker." In comparison to Figure 2a, we see that when a user exhibiting this role chooses to comment, they tend to do so on their own question. This may indicate engagement with users exhibiting the "clarifier" role (see Figure 2e) to improve the question's quality before obtaining an answer.



(e) A "clarifier" role. The majority of this user's activity is centered on commenting behavior, and is predominately engaging in discussion on others' questions. This is likely a result of this type of user engaging with others exhibiting the "careful asker" role (see Figure 2d) in order to clarify the question before providing an answer.

Figure 2: The role distributions discovered by our MDMM behavior model fit to 161 StackExchange communities. The labels given to these roles reflect our own interpretation of the role and are given here to make disambiguating the roles easier in the text. The MDMM behavior model can uncover subtle distinctions in asking behavior (see Figure 2a vs Figure 2d) and answering behavior (see Figure 2c vs Figure 2e).

iors in a community, i.e., a behavior profile. This profile provides a representation of a community that can be further analyzed, as we will discuss in this section.

We start with the following question: are there systematic differences in role proportions between groups of communities in our dataset? To answer this question, we grouped each community in the StackExchange dataset using the taxonomy provided by StackExchange itself[5]: (1) Technology, (2) Culture/Recreation, (3) Life/Arts, (4) Science, (5) Professional, and (6) Business. To allow for a "warm-up" period for the community and to eliminate the issue of noisy proportion vectors arising due to data sparsity during community launch, we discard the first 12 months of role proportion data for each community. We then only consider communities that have at least 12 months of data beyond that warm-up period to allow for computing an average proportion vector to represent the community over at least one year. After filtering, the "Professional" and "Business" groups have only five and four communities, respectively, so we consider only the four larger groups. "Technology" had 52 communities, "Culture/Recreation" had 36, "Life/Arts" had 20, and "Science" had 17. We show the group memberships in Table 3.

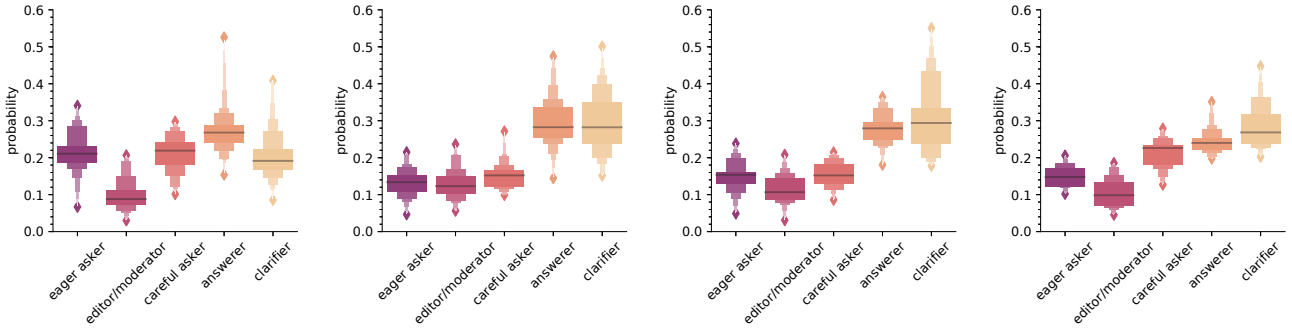
These four groups' role proportions are visualized in Figure 3. Visually, we can see a number of differences. First, the "eager asker" role is more prominent in the "Technology" group than all three others. Both the "Technology" and "Science" groups have higher prominence of the "careful asker" role when compared against "Culture/Recreation" and "Life/Arts". We can also see that the "clarifier" role is diminished in the "Technology" compared to the others.

There are also notable commonalities between groups. The "Culture/Recreation" and "Life/Arts" groups are quite similar across nearly all of the roles. The "editor/moderator" role prevalence is similar across all of the groups (with only a slight increase present for the "Culture/Recreation" group). "Answerer" prevalence is similar across all of the groups (where the reduction in variance in "Life/Arts" and

[5]https://stackexchange.com/sites

Table 3: Communities belonging to each of the four groups we consider from StackExchange's own taxonomy.

Group	Members
Technology	android, apple, arduino, askubuntu, bitcoin, blender, codegolf, codereview, craftcms, crypto, datascience, dba, drupal, dsp, ebooks, electronics, emacs, expressionengine, gamedev, gis, ja.stackoverflow, joomla, magento, mathematica, networkengineering, opendata, programmers, pt.stackoverflow, raspberrypi, reverseengineering, robotics, ru.stackoverflow, salesforce, security, serverfault, sharepoint, softwarerecs, sound, space, sqa, stackapps, stackoverflow, superuser, tex, tor, tridion, unix, ux, webapps, webmasters, windowsphone, wordpress
Culture/Recreation	anime, beer, bicycles, boardgames, bricks, buddhism, chess, chinese, christianity, ell, english, french, gaming, german, ham, hermeneutics, hinduism, history, homebrew, islam, italian, japanese, judaism, martialarts, mechanics, outdoors, poker, politics, puzzling, rpg, rus, russian, skeptics, spanish, sports, travel
Life/Arts	academia, avp, cooking, diy, expatriates, fitness, gardening, genealogy, graphicdesign, lifehacks, money, movies, music, parenting, pets, photo, productivity, scifi, sustainability, worldbuilding
Science	astronomy, biology, chemistry, cogsci, cs, cstheory, earthscience, economics, hsm, linguistics, math, matheducators, mathoverflow.net, philosophy, physics, scicomp, stats



(a) Technology (b) Culture/Recreation (c) Life/Arts (d) Science

Figure 3: Letter-value plots of the role proportion vectors for the four largest StackExchange groups after filtering communities with less than 12 months of data after filtering a start-up period of 12 months.

"Science" likely attributable to there being fewer communities in those groups).

To quantify the statistical significance of the above observations, we use a Kruskal-Wallis H test (Kruskal and Wallis 1952) to perform a one-way ANOVA test to determine the existence of a difference between a single role proportion across all four groups, for each role proportion. Then, if a statistically significant difference between the groups is reported, we use a post-hoc Conover-Iman test (Conover and Iman 1979) to determine which of the groups exhibit statistically significant differences in that role proportion. To correct for multiple testing in both cases, we use the Holm-Bonferroni method (Holm 1979) to correct the p-values. We report our findings in Table 4. On the whole, we see that the "Technology" group differs strongly from the other three groups in terms of its proportion of "eager asker" (where it is higher) and "clarifier" roles (where it is lower). We also see that the "careful asker" role is more prominent in the communities from the "Technology" and "Science" groups and less prominent in the "Culture/Recreation" and "Life/Arts" groups. This suggests that the more technical communities in "Technology" and "Science" require more discussion around questions than the less technical communities of "Culture/Recreation" and "Life/Arts".

Thus, we have demonstrated the utility of using the MDMM behavior model for understanding differences in user behavior across communities. This is easily facilitated because it learns a role proportion vector $\theta_{1:N}$ that, by design, can be readily interpreted in the context of the discovered roles $\phi_{1:K}$.

4.4 Behavior Compositions and their Relationship to Community Success

As another example of what one can learn by studying the community role compositions that can be discovered by the MDMM user behavior model, we now ask the following question: how does the proportion of roles within a community relate to its success? In order to explore this, we first need to be able to define what we mean by "success" in a CQA community. We have taken a content-focused approach to understanding behavior, so we also choose to define the success of a community in terms of its content generation. Borrowing from Dev et al. (2018), we have the following metrics: (1) the ratio of the number of answers N_a to the number of questions N_q, which is a reflection of the ability of a community to cope with question load; (2) the percent-

Table 4: Statistical significance tests for differences in role proportions across the four groups. All p-values are adjusted using the Holm-Bonferroni method. Shown are only those tests that are statistically significant at a threshold of 0.05. We notice strongly significant differences ($p < 1 \times 10^{-5}$) in role proportions for the "eager asker", "careful asker", and "clarifier" roles.

Role	p-value	Group Pair	p-value
eag. ask.	3.87×10^{-11}	cult. tech.	1.49×10^{-14}
		life vs. tech.	5.41×10^{-7}
		sci. vs. tech.	6.63×10^{-7}
edit/mod.	1.10×10^{-2}	cult. vs. tech.	2.40×10^{-3}
care. ask.	7.53×10^{-9}	cult. vs. sci.	3.00×10^{-6}
		cult. vs. tech.	3.41×10^{-9}
		life vs. sci.	5.80×10^{-5}
		life vs. tech.	3.07×10^{-6}
answerer	1.10×10^{-2}	cult. vs. sci.	4.44×10^{-3}
clarifier	4.41×10^{-8}	cult. vs. tech.	5.22×10^{-8}
		life vs. tech	1.69×10^{-6}
		sci. vs. tech	2.30×10^{-5}

age of questions that receive an answer; (3) the percentage of questions that receive an "accepted" answer[6], which reflects the community's ability to provide high-quality answers to new questions; and finally (4) the average time before the arrival of the first answer[7], which measures the timeliness of the community's answering capabilities.

Each of these metrics can be computed for each monthly snapshot of a community (by considering the questions that are asked within that time period). Then, we can average the value for a metric across all of the months of a community to obtain an overall score for that metric for that community. We again only consider the communities that, after dropping 12 months of "warm-up" period data, have at least 12 months of data.

The results are visualized in Figure 4. While differences in these metrics are small, they are statistically significant (see Table 5). In particular, we notice that the "Culture/Recreation" and "Life/Arts" groups have a higher ratio of answers to questions (Figure 4a) and a higher fraction of answered questions (Figure 4b) when compared to the "Science" and "Technology" groups. These same pairs exhibit statistically significantly different proportions of the "careful asker" role.

This provides an interesting insight: groups of communities that have a higher propensity for the "careful asker" role exhibited *lower health metrics* across multiple measures. In fact, every pair of groups that exhibited a statistically significant difference in this role proportion also had statistically

[6]On StackExchange, the original poster of a question can designate one of the answers provided as being "correct" by "accepting" that answer.

[7]We compute this only for questions that did receive an answer.

Table 5: Statistical significance tests for differences in health metrics across the four groups. All p-values are adjusted using the Holm-Bonferroni method. Shown are only those tests that are statistically significant at a threshold of 0.05. We note that, with a single exception ("Science" vs "Technology"), when there is a statistically significant difference in role proportions, there is a statistically significant difference in at least one of the four health metrics we explore. Similarly, groups that do not have different role proportions ("Culture/Recreation" and "Life/Arts") do not have significant differences in health metrics.

Metric	p-value	Group Pair	p-value
N_a/N_q	7.08×10^{-7}	cult. vs. sci.	4.26×10^{-5}
		cult. vs. tech.	3.51×10^{-6}
		life vs. sci.	1.20×10^{-4}
		life vs. tech.	6.50×10^{-5}
% ans.	6.34×10^{-5}	cult. vs. sci.	7.44×10^{-5}
		cult. vs. tech.	4.12×10^{-4}
		life. vs. sci.	3.16×10^{-3}
		life. vs. tech.	3.36×10^{-2}
% acc. ans.	1.08×10^{-2}	cult. vs. sci.	6.68×10^{-3}
Resp. time	1.08×10^{-2}	cult. vs. sci.	2.31×10^{-2}
		cult. vs. tech.	3.34×10^{-2}

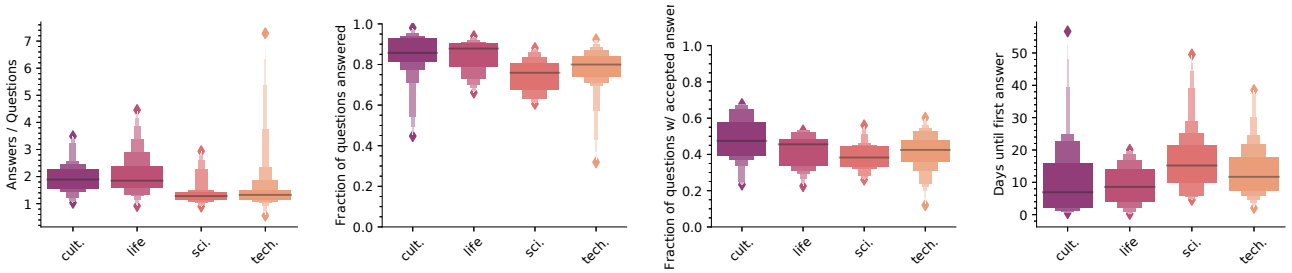
significant differences present in at least two metrics (with one pair with three and another with four). While we cannot say whether this correlation is causal, this opens the door for more studies into impact of the "careful asker" profile on community health—a question we could not have raised without first having a tool like the MDMM behavior model to aid our efforts to understand user behavior.

Furthermore, notice that groups that do *not* exhibit differences in their behavior profiles (namely "Culture/Recreation" and "Life/Arts") also do not exhibit differences in any of our four health metrics.

4.5 Evolution of Behavior Composition

The questions we have explored so far have focused mainly on static snapshots of the CQA communities in our dataset. However, these communities do not exist in a vacuum—they continually evolve over time as they acquire new users and address new topics. How can we understand how community behavior changes over time as these communities grow and evolve? Here, we explore one potential solution using the MDMM behavior model as yet another example application.

Because we segmented the user sessions by month for each community, we have a role proportion associated with each (community, month) pair. With this information in hand, we can then plot a collection of time-series for each community by considering the role proportions for each individual role over the life of the community. This plot can allow us to understand how role proportions fluctuate as the community evolves. In Figure 5, we show the evolution of the top three oldest communities belonging to the "Tech-



(a) N_a/N_q (b) Frac. of N_q with ans. (c) Frac. of N_q with accepted ans. (d) Avg. days until first answer

Figure 4: Health metrics for each of the four groups of StackExchanges considered in Section 4.3. Differences are small but statistically significant (see Table 5). N_a/N_q is higher for "Culture/Recreation" and "Life/Arts" than for "Science" and "Technology". Similarly, "Culture/Recreation" and "Life/Arts" enjoy a higher fraction of answered questions compared to "Science" and "Technology", and also have faster average response times (though only "Culture/Recreation" statistically significantly so). "Culture/Recreation" also has a higher fraction of questions with an accepted answer compared to the other three groups.

nology" and "Culture/Recreation" groups, respectively. We start plotting the time series at the month when the community first has at least 100 browsing sessions.

We can see a few trends occurring. First, we can see a common trend in Figure 5a–5c, where the proportions for the "eager asker" role grow, reach a peak within the first quarter or so of the community's life, and then begin a steady decline over time. We also notice that the "careful asker" and "clarifier" roles tend to increase steadily over time, nearly in tandem. Second, we can see in Figure 5d–5f that the role proportions tend to be more consistent over time for members of the "Culture/Recreation" group than for "Technology". Note, however, that the exact composition that is remaining stable varies between the communities. That is to say, communities in "Culture/Recreation" appear to be more stable relative to themselves over time, but exhibit variation in what that stability looks like.

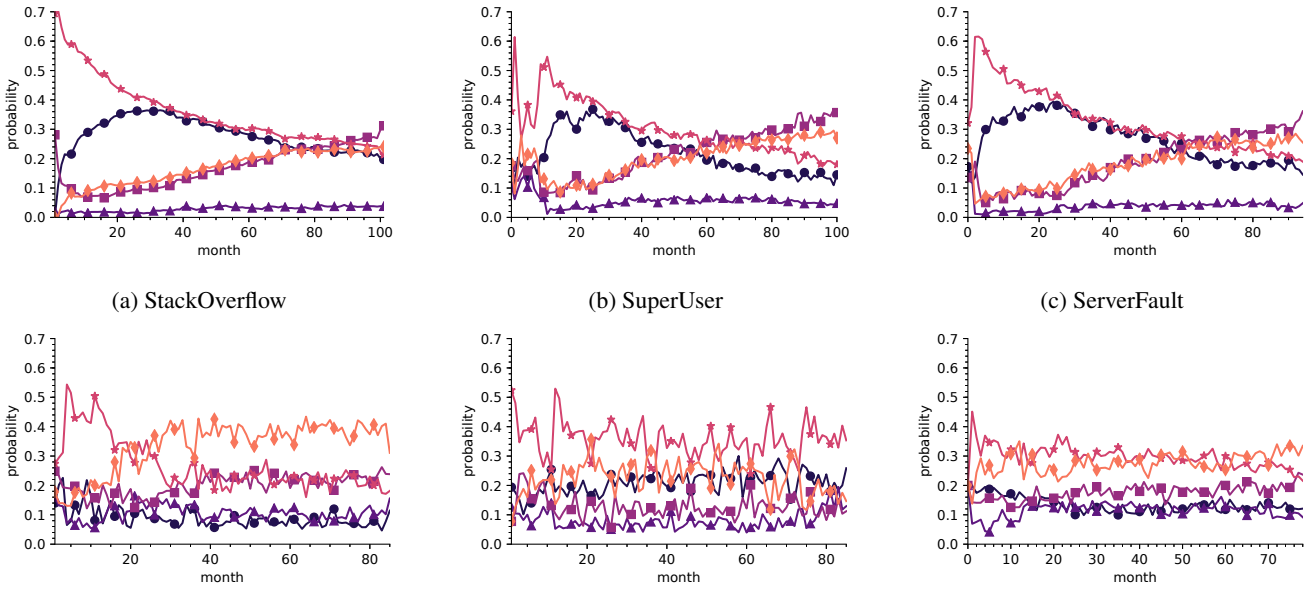
Why does this behavior shift happen in "Technology" while "Culture/Recreation" communities remain more stable? While we cannot yet provide an answer to this question, we note that without first being able to see that this kind of behavior evolution is even taking place (which requires a model like our MDMM behavior model), we could not even begin to ask such a question. This shows that the MDMM behavior model opens new interesting research directions in understanding user behavior in ways we were not able to before.

5 Discussion and Limitations

The goal of this work is to contribute a new and general tool for role discovery and analysis of community role compositions. There are two key ideas in the design of the proposed model. The first is the formalization of a *shared* set of user roles, distributions over user actions, across communities. This is an expressive representation of a user role as the distribution can vary to capture subtle differences between user roles while also allowing us to discover user roles empirically from the data using sound statistical principles. The second is the direct modeling of the composition of user roles in a CQA community with another distribution over the user roles. This second distribution provides a general and flexible way to model variations in the composition of user roles that may exist in different communities, and again allows us to use statistical inference to discover each community's role composition.

The use of a generative model over user actions to discover user roles and community role compositions is advantageous in that it allows the model to be very general and applied in a variety of different analysis scenarios without requiring hand-crafted features to be defined in order to describe user roles. On the other hand, the use of a generative model is not without some cost. Because statistical inference of such a model is intractable, we must resort to approximate posterior inference methods. In this paper, we have used Gibbs sampling to approximate the posterior, but this comes with some risk—it is difficult to determine whether the sampler has actually converged to the true posterior, despite there being a theoretical guarantee that it will do so given enough time. Had we instead opted for a different inference method like variational inference which instead optimizes a variational lower bound, we trade the convergence question for a question about the quality of the solution found by the optimization because the variational lower bound is highly non-convex. In practice, we can attempt to mitigate these concerns via multiple runs of the sampler (or multiple randomly initialized optimizations for variational inference)—we found multiple runs of the model all converged to nearly identical solutions.

Because the model does not impose an action set upon the user, they are free to specify a different action set for different analysis purposes. This again makes the model quite flexible, but also requires some up front work to define an appropriate action set for the model. Feeding the model with less meaningful actions can lead to the output of less meaningful role patterns. Fortunately, in the case of CQA communities, defining an action set based on the content hierarchy and content ownership semantics is a reasonable choice that should lead to interpretable roles as demonstrated here for StackExchange. However, a user *does* need to manually interpret the role distributions $\phi_{1:K}$ discovered by the model.



(a) StackOverflow (b) SuperUser (c) ServerFault

(d) Judaism (e) Homebrew (f) Gaming

Figure 5: Role proportions over time for the three oldest communities belonging to the "Technology" and "Culture/Recreation" groups. (a)–(c) belong to the "Technology" group, and (d)–(f) belong to the "Culture/Recreation" group. We can see a common trend in (a)–(c) where the proportion of the "eager asker" role grows until it peaks, and then declines as the community ages. The "clarifier" and "careful asker" roles increase over time, almost in tandem in this group. However, in (d)–(f) we see that that communities belonging to "Culture/Recreation" tend to have role proportions that remain more consistent over time (in that they do not demonstrate long-term trends.)

Finally, our model makes a strong assumption that a user only performs one role in a given session. While this assumption is valid in most cases, there are situations where users potentially perform more than one role in a given browsing session. In these cases, the model will incorrectly conflate these two roles and this will contribute some "noise actions" to that role.

6 Conclusion and Future Work

Computational analysis of user roles on CQA platforms is important not only for the understanding of users in such a new social network environment, but also for improving their efficiency and utility. To this end, we proposed a general probabilistic model for discovering and analyzing action-based roles on CQA platforms. The generative model assumes that the observed user actions in a single session are samples drawn from the same, but unknown, action distribution (the role). Individual communities are modeled as mixtures over these role distributions, allowing for cross-community analysis. Through a comprehensive experiment on all 161 non-meta communities on the StackExchange CQA platform, we demonstrated that our model is indeed useful for understanding user behavior on these platforms. We were able to show interesting distinctions in asking and answering behavior on the platform are captured through our roles, that different groups of communities exhibit statistically significant differences in role composition, and those communities also exhibit statistically significant differences in a variety of health measures. Finally, we were also able to uncover two clear and distinct trends of role compositions over time between the "Technology" and "Culture/Recreation" groups on StackExchange.

The proposed model is very general and does not require labeled data for training. It can thus be applied to analyze any CQA platform immediately. Since the definition of actions is outside the model, analysts can vary the granularity of actions as needed; this flexibility allows for multi-resolution analysis of user actions, behavior, and roles. An interesting future work is to fully exploit this flexibility to further analyze roles with even more refined actions on CQA platforms as well as to apply the model to other social networks. Another interesting future direction is to develop tools based on this model for monitoring the "well-being" of those CQA platforms and helping the community managers to improve the utility and efficiency of a community so as to maximize the utility of all the CQA communities.

Acknowledgments

This material is based upon work supported by the NSF GRFP under Grant Number DGE-1144245, and by the NSF Research Program under Grant Number IIS-1629161.

References

Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge sharing and yahoo answers: Everyone

knows something. In *Proc. WWW*, WWW '08, 665–674.

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2012. Discovering value from community activity on focused question answering sites: A case study of stack overflow. In *Proc. KDD*, KDD '12, 850–858.

Barash, V. D.; Smith, M.; Getoor, L.; and Welser, H. T. 2009. Distinguishing knowledge vs social capital in social media with roles and context. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*.

Benevenuto, F.; Rodrigues, T.; Cha, M.; and Almeida, V. 2009. Characterizing user behavior in online social networks. In *Proc. IMC*, IMC '09, 49–62.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Chan, J.; Hayes, C.; and Daly, E. M. 2010. Decomposing discussion forums and boards using user roles. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 215–218.

Conover, W. J., and Iman, R. L. 1979. On multiple-comparisons procedures. Technical report, Los Alamos Scientific Laboratory.

Dev, H.; Geigle, C.; Hu, Q.; Zheng, J.; and Sundaram, H. 2018. The size conundrum: Why online knowledge markets can fail at scale. In *Proceedings of WWW 2018: The Web Conference*, 65–75. New York, NY, USA: ACM.

Fisher, D.; Smith, M.; and Welser, H. T. 2006. You are who you talk to: Detecting roles in usenet newsgroups. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences - Volume 03*, HICSS '06, 59.2–. Washington, DC, USA: IEEE Computer Society.

Furtado, A.; Andrade, N.; Oliveira, N.; and Brasileiro, F. 2013. Contributor profiles, their dynamics, and their importance in five q&a sites. In *Proc. CSCW*, CSCW '13, 1237–1252.

Geigle, C., and Zhai, C. 2017. Modeling mooc student behavior with two-layer hidden markov models. *Journal of Educational Data Mining* 9(1):1–24.

Gündüz, c., and Özsu, M. T. 2003. A web page prediction model based on click-stream tree representation of user behavior. In *Proc. KDD*, KDD '03, 535–540.

Han, Y., and Tang, J. 2015. Probabilistic community and role model for social networks. In *Proc. KDD*, KDD '15, 407–416.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *Proc. UAI*, UAI'99, 289–296.

Holm, S. 1979. A simple sequentially rejective multiple test proceedure. *Scandinavian Journal of Statistics* 6(2):65–70.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Kruskal, W. H., and Wallis, W. A. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47(260):583–621.

Kullback, S. 1959. *Information Theory and Statistics*. John Wiley & Sons.

Lu, L.; Dunham, M.; and Meng, Y. 2006. Mining significant usage patterns from clickstream data. In *Proc. WebKDD*, WebKDD'05, 1–17.

Maddison, C. J.; Tarlow, D.; and Minka, T. 2014. A* sampling. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 3086–3094.

Mamykina, L.; Manoim, B.; Mittal, M.; Hripcsak, G.; and Hartmann, B. 2011. Design lessons from the fastest q&a site in the west. In *Proc. CHI*, CHI '11, 2857–2866.

Manavoglu, E.; Pavlov, D.; and Giles, C. L. 2003. Probabilistic user behavior models. In *ICDM*, 203–210.

McCallum, A.; Wang, X.; and Corrada-Emmanuel, A. 2007. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Int. Res.* 30(1):249–272.

Nam, K. K.; Ackerman, M. S.; and Adamic, L. A. 2009. Questions in, knowledge in?: A study of naver's question answering community. In *Proc. CHI*, CHI '09, 779–788.

Nigam, K.; Mccallum, A. K.; Thrun, S.; and Mitchell, T. 2000. Text classification from labeled and unlabeled documents using em. *Machine Learning* 39(2):103–134.

Qiu, M.; Zhu, F.; and Jiang, J. 2013. It is not just what we say, but how we say them: Lda-based behavior-topic model. In *Proc. SDM*, SDM '13, 794–802.

Sadagopan, N., and Li, J. 2008. Characterizing typical and atypical user sessions in clickstreams. In *Proc. WWW*, WWW '08, 885–894.

Su, Q., and Chen, L. 2015. A method for discovering clusters of e-commerce interest patterns using click-stream data. *Electron. Commer. Rec. Appl.* 14(1):1–13.

Wang, G.; Zhang, X.; Tang, S.; Zheng, H.; and Zhao, B. Y. 2016. Unsupervised clickstream clustering for user behavior analysis. In *Proc. CHI*, CHI '16, 225–236.

Wang, S.; Lo, D.; and Jiang, L. 2013. An empirical study on developer interactions in stackoverflow. In *Proc. SAC*, SAC '13, 1019–1024.

Welser, H. T.; Gleave, E.; Fisher, D.; and Smith, M. 2007. Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure* 8:1–31.

White, A.; Chan, J.; Hayes, C.; and Murphy, T. B. 2012. Mixed membership models for exploring user roles in online fora. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 599–602.

Wu, L.; Baggio, J. A.; and Janssen, M. A. 2016. The role of diverse strategies in sustainable knowledge production. *PLOS ONE* 11(3):1–13.

Xu, Z.; Zhang, Y.; Wu, Y.; and Yang, Q. 2012. Modeling user posting behavior on social media. In *Proc. SIGIR*, SIGIR '12, 545–554.

Yin, J., and Wang, J. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proc. KDD*, KDD '14, 233–242.