Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms

Isaac Waller
University of Toronto
walleris@cs.toronto.edu

Ashton Anderson
University of Toronto
ashton@cs.toronto.edu

ABSTRACT

In many online platforms, people must choose how broadly to allocate their energy. Should one concentrate on a narrow area of focus, and become a specialist, or apply oneself more broadly, and become a generalist? In this work, we propose a principled measure of how generalist or specialist a user is, and study behavior in online platforms through this lens. To do this, we construct highly accurate *community embeddings* that represent communities in a high-dimensional space. We develop sets of community analogies and use them to optimize our embeddings so that they encode community relationships extremely well. Based on these embeddings, we introduce a natural measure of activity diversity, the GS-score.

Applying our embedding-based measure to online platforms, we observe a broad spectrum of user activity styles, from extreme specialists to extreme generalists, in both community membership on Reddit and programming contributions on GitHub. We find that activity diversity is related to many important phenomena of user behavior. For example, specialists are much more likely to stay in communities they contribute to, but generalists are much more likely to remain on platforms as a whole. We also find that generalists engage with significantly more diverse sets of users than specialists do. Furthermore, our methodology leads to a simple algorithm for community recommendation, matching state-of-the-art methods like collaborative filtering. Our methods and results introduce an important new dimension of online user behavior and shed light on many aspects of online platform use.

KEYWORDS

generalist and specialists, community embeddings, activity diversity, community recommendation

ACM Reference Format:
Isaac Waller and Ashton Anderson. 2019. Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313729

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313729

1 INTRODUCTION

> The fox knows many things; but the hedgehog knows one big thing.
>
> *Archilochus, 7th-century BC*

In many domains of human endeavor, people must choose how broadly to allocate their energy. Should one be a *specialist*, and concentrate in depth on a narrow area of focus, or be a *generalist*, and apply oneself more broadly? In the past century, there has been a steady movement towards specialization in professional domains, including business, medicine, and academia [5, 12, 22], while in the past decade, there has been a rise of generalists, especially in entrepreneurship and in the tech industry [6, 21].

The trade-off between breadth and depth of activity is also a fundamental characteristic of user behavior in online platforms. A user can concentrate on a narrow range of activities and reap the rewards of specializing in some area of the platform, or one can maintain a more diverse set of interests at the cost of not engaging as deeply with them. Whether the choice is explicitly made or not, the scope of a user's activity ranges between being narrow and broad. There exists a continuum of possibilities between extreme specialists on one side and extreme generalists on the other, and in general users will be somewhere in the middle.

Activity Diversity in Online Platforms. The question of how diverse users' activities are illuminates a number of important issues. First, the extent to which users are generalists or specialists indicates how connected an online platform is on a global level. If most users lean towards being specialists, then the platform will tend to consist of isolated communities with little interaction between them, whereas if most users lean towards being generalists, then the platform will tend to be more cohesive and support broader engagement between members of different communities. Second, users' activity diversity styles can be indicative of the underlying incentives they face. As in medicine and academia, where the prevailing incentives once rewarded generalists but now typically steer professionals towards specialization, a preponderance of specialists or generalists in online platforms could reveal how users perceive their incentive structures. Third, the relation between breadth and depth of activity and resulting success and longevity has been the source of rich debates in other contexts [2, 18]. The study of how this plays out in online platforms could contribute to this discussion.

Despite the fundamental nature of the activity diversity problem, and its relative simplicity to state, it is difficult to study rigorously for several reasons. In the simplest form of the problem, we are given a set of activities supported by an online platform, and empirical data on what activities individual users choose to pursue. The main obstacles in measuring which users are generalists and which are specialists are twofold: first, coming up with measures of similarity between all pairs of activities, and second, ensuring

that these similarities are consistent with each other. In prior work, researchers have mainly used entropy as a measure of activity diversity [7]. Entropy is very sensible in many scenarios, and is sensitive to both the number and distribution of activities. However, it is agnostic to the similarities or differences between activities. Thus, a user whose activities are all very similar to each other would be treated the same as a user with the same distribution of engagement across activities that are very different from each other, contradicting the intuition that the former is more of a specialist.

The Present Work: Community Embeddings and the GS-Score. In this work, we present a methodology to study generalists and specialists in online platforms, and we analyze, characterize, and predict user behavior on Reddit and GitHub through this lens. On Reddit, a discussion-based content aggregation platform, we are interested in how users participate in various communities (*subreddits*) as defined by commenting on posts in those communities; and on GitHub, a social coding platform, we are interested in how users interact with (contribute, star, or watch) various code repositories.

To solve the problem of deriving high-fidelity, consistent similarities between communities, we develop and considerably extend *community embeddings*. Following previous work [11, 15], we adapt word2vec, a word embedding algorithm that assigns a vector to each word in a training corpus such that relationships between words are preserved, to assign a vector to each community in our data such that relationships between communities are preserved. In word embeddings, analogical reasoning can surprisingly be done with simple vector arithmetic. Our first contribution is developing a set of ground-truth community analogies for both Reddit and GitHub that can be used to optimize the embedding algorithm hyperparameters and vastly enhance the quality of the resulting community embeddings. In our final embeddings, analogy questions like `toronto : torontoraptors :: chicago : ___`? are answered correctly (`chicagobulls`) with extremely high accuracy.

Using our community embeddings, we are able to quantify the similarity between any two communities as the cosine similarity between their respective vector representations. Furthermore, these similarities are mutually consistent, as a result of being derived from the geometry of the same vector space. We represent a user's activities as a distribution over points in this vector space. Intuitively, a specialist's activity is concentrated in the space, and a generalist's activity is diffuse in the space (see Figure 1 for a schematic depiction). We calculate a user's *center of mass* as being the centroid of the communities they contribute to, and we define the *Generalist-Specialist score* of a user's activity to be the average cosine similarity between their communities and their center of mass, weighted by activity. The GS-score is minimized by specialists who contribute to a tight cluster of communities and maximized by generalists who contribute to communities that are far apart. We developed an online quiz to elicit human judgments about generalist and specialist users and found that they correlate very well with the GS-score. This principled measure of activity diversity in online platforms, using high-fidelity and validated community similarities calculated from community embeddings, is a key contribution of our work.

Overview of Results. Our results fall into two categories, according to our two main units of analysis: the user and the community.

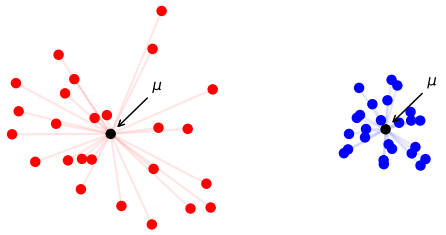


Figure 1: A schematic depicting the vector representations of communities contributed to by a generalist (left) and a specialist (right). The generalist's communities are spread out, and the specialist's communities are clustered together.

We first focus on users. Applying the GS-score to Reddit and GitHub users, we observe a great diversity of user activity styles, ranging from extreme generalists to extreme specialists, at every level of activity. Compared to a random baseline, there are many more specialists than expected. On Reddit, we analyze how users' comments are voted on, and find that specialists do about 20% better in a controlled analysis. We then turn to studying user longevity, both in individual communities and on the platform as a whole. We find that specialists are much more likely than generalists to convert into long-term members of a particular community, suggesting that specialists are more selective. However, we also observe that generalists are much more likely to remain on the platform, with the effect being particularly stark on Reddit. These results have interesting implications for community managers and platform designers; on the one hand a community manager should probably seek out specialists to join their community, but on the other hand platform designers should try to be amenable to generalists and encourage exploration. Finally, we measure the diversity of the subpopulations users interact with, and find that specialists interact with significantly less diverse sets of users than generalists do.

We then turn our attention to communities, and make the observation that we can define the activity diversity of a *community* to be the average activity diversity of its users. We find this *community activity diversity* score to be remarkably stable, remaining relatively constant for at least three years, even when the underlying user base turns over significantly. We also analyze which users make up the "elites" in a community and find that they skew generalist.

Finally, we apply our measure in a prediction framework and focus on two concrete, important tasks. The first is community recommendation, in which we predict which communities users will join. We develop a simple algorithm to recommend communities based on their proximity to a user's center of mass and demonstrate that it performs at least as well as state-of-the-art collaborative filtering. Furthermore, we find that specialists are much more predictable than generalists. The second problem is user retention, in which we predict if users will stay on the platform. We show that using activity diversity alone performs very well.

The rest of the paper is laid out as follows. In Section 2, we discuss the datasets, develop community embeddings, and introduce our metric of activity diversity, the GS-score; in Section 3 we analyze behavior in Reddit and GitHub through the lens of activity diversity; in Section 4 we perform our prediction tasks; in Section 5 we outline related work; and in Section 6 we discuss our work and conclude.

2 METHODOLOGY

The main goal of our work is to precisely define a measure of activity diversity that smoothly interpolates between generalists and specialists, and apply it to the study of user behavior in online platforms. In this Section, we describe our datasets, develop our methodology to construct high-fidelity community embeddings, validate these embeddings as being accurate representations of the platforms we study, and introduce our measure of activity diversity.

2.1 Data

First we introduce the online platform datasets we use throughout the rest of the paper. We study two online platforms in depth: Reddit, a discussion-based content aggregation platform, and GitHub, a collaborative coding platform.

Reddit. Our main dataset is derived from Reddit, an online platform where users can share, view, and discuss content from around the Web. Every post is either a piece of free-standing text or a URL, and there is a discussion page where users can discuss the post. Reddit is subdivided into communities called *subreddits* that are organized around some theme; for example `r/gardening` is about gardening and `r/learnprogramming` is for people learning how to program. We are interested in user participating in these communities by *commenting* in them, i.e. by participating in discussions on discussion pages. Comments can also be voted on by other users, and each has a vote score associated with it.

Our Reddit dataset is comprised of all Reddit user comments in 2017, totalling 900M comments authored by 11.4M distinct users in 232K subreddits. We restrict our attention to the top 10,000 subreddits by activity, which account for 96.8% of all comments. The data we used is publicly available and can be downloaded from `pushshift.io` [3].

GitHub. We supplement our Reddit analyses with a dataset of GitHub activities. GitHub is an online platform where users can collaborate on coding projects that are divided into *repositories* of self-contained projects. Popular repositories frequently support a vibrant community of contributors. We treat repositories as communities, and consider adding code and bookmarking to be user contributions to these communities.

Our GitHub dataset is comprised of all GitHub commits, pull requests, forks, watches, and stars in 2017, totalling 413M actions by 8.3M distinct users in 26M repositories. Since the vast majority of repositories have very few contributors (usually 1), we restrict our attention to the top 40,000 communities by number of stars. The data we used is publicly available from the GH Archive and can be downloaded from `gharchive.org`.

2.2 Community Embeddings

As mentioned in the Introduction, a chief difficulty in formulating a measure of activity diversity is coming up with accurate similarities between all pairs of communities. Without these, it is hard to distinguish between a user who is active in two very related communities and a user who is active in two very different communities, despite the fact that intuitively the former should be treated as more specialist than the latter. The most commonly used metric for capturing user diversity among activities is entropy, which is agnostic to community similarities, and therefore treats these two users as being equally diverse in their activities.

Table 1: Example community analogies from our collection.

a	a'	b	b'	
UCSD	sandiego	georgetown	washingtondc	university to city
Harvard	CambridgeMA	mcgill	montreal	
Northwestern	evanston	UCL	london	
vancouver	canucks	Detroit	DetroitRedWings	city to team
toronto	torontoraptors	chicago	chicagobulls	
Seattle	Seahawks	oakland	oaklandraiders	
SFGiants	baseball	49ers	nfl	team to sport
lakers	nba	Dodgers	baseball	

To solve this problem, we construct high-fidelity *community embeddings*. Drawing inspiration from the major successes achieved by applying word embeddings to various natural language processing tasks, and following recent work [11, 15], we adapt the word2vec word embedding algorithm to apply to user-community interaction data to generate community embeddings. Treating communities as "words" and users who comment in them as "contexts", we embed communities into a high-dimensional vector space. Two communities are close together if users frequently comment in them both, and are far apart if they are rarely commented in by the same users. We use the skip-gram model with negative sampling and train over all pairs (c_i, u_j) of user u_j commenting in community c_i.

Using the same hyperparameters as those used in the original word2vec model, we generated an initial community embedding. It seems to work reasonably well — for example, the nearest communities to `r/Fishing` (those with highest cosine similarity with `r/Fishing`) are focused on other outdoor pursuits, e.g. `r/discgolf`. However, it is unclear how to objectively evaluate the fidelity of this embedding.

This problem of objectively measuring embedding quality also arises with word embeddings. There, researchers and practitioners solve this problem by making use of the surprising fact that embeddings preserve semantic relationships. For example, in a well-tuned word embedding, the analogy `man : king :: woman : ___?` can be correctly solved with `queen` by performing vector arithmetic on the corresponding word vectors ($\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}} \approx \overrightarrow{\text{queen}}$). By generating a set of word analogies and then measuring how many of them a word embedding can solve by performing vector arithmetic, one can measure the quality of a word embedding.

Community analogies. To evaluate our community embeddings, we adapt this methodology to our community setting and develop large sets of ground-truth *community analogies*, which can be used to objectively evaluate the quality of a community embedding. Each analogy is a tuple (a, a', b, b') of communities such that $a : a' :: b : b'$, and it should be the case that $\overrightarrow{\text{a'}} - \overrightarrow{\text{a}} + \overrightarrow{\text{b}} \approx \overrightarrow{\text{b'}}$. We found three different classes of objective semantic relationships on Reddit communities that we can generate analogies with: university → city, sports team → city, and sports team → sport. For example, for the first class of analogies, university communities (e.g. `r/Harvard`) should all have roughly the same vector relationship with the communities of the city they are located in (e.g. `r/CambridgeMA`). Thus, it should be the case that e.g., $\overrightarrow{\texttt{r/Harvard}} - \overrightarrow{\texttt{r/CambridgeMA}} \approx \overrightarrow{\texttt{r/mcgill}} - \overrightarrow{\texttt{r/montreal}}$. Example analogies of each type are shown

in Table 1. For each class of analogies, we generated a list of pairs of communities with the same semantic relationship, then took the Cartesian product of these pairs to construct all possible analogies for this class. Across our three classes, we constructed 4,392 analogies in total[1].

We can now validate our embedding. For each analogy, we rank all vector representations of communities in ascending order of distance to the point $\overrightarrow{a'} - \overrightarrow{a} + \overrightarrow{b}$. The accuracy metric Precision@k measures the fraction of analogies for which $\overrightarrow{b'}$ is among the k closest communities, and the special case Precision@1 measures what fraction of analogies are perfect, i.e. $\overrightarrow{b'}$ is the closest community to the point $\overrightarrow{a'} - \overrightarrow{a} + \overrightarrow{b}$. Our initial embedding scores 30% Precision@1 on our set of 4,392 validation analogies.

Hyperparameter search. To optimize our embedding, we conducted a sweep over the space of word2vec parameters, generating a separate community embedding for every combination of parameter values and evaluating all of these embeddings with our analogy sets. The parameters we varied are the learning rate α, the dimensionality of the vector space, the number of negative samples, and the sampling rate. How analogy performance varies with parameters is shown in Figure 2. Optimizing the parameters in this way drastically improves the resulting community embedding. Our final embedding scores 72% Precision@1 and 96% Precision@5 on our analogy test, a massive improvement over our initial embedding scores. We show a two-dimensional projection of our final embedding in Figure 3 centered on `r/Fishing`.

Optimizing the embedding with analogies has achieved a difference in kind in the fidelity of the resulting embedding; now all of the closest communities are extremely related (e.g. `r/bassfishing`, `r/kayakfishing`, and `r/flyfishing` are the top three closest communities). We emphasize that the fact this is even possible was not a priori obvious. Community embeddings preserve semantic relationships between communities much as word embeddings do with words. Furthermore, similarly to word embeddings, optimizing community embeddings with analogies drawn from a small number of classes of semantic relationships seems to generally align the entire vector space, even for relationships we didn't train on. The vast difference in quality between the neighborhoods of `r/Fishing` in the optimized and unoptimized embeddings is typical. Also, we observe that other types of analogies suddenly work, even though they weren't explicitly trained; for example, the analogy $\overrightarrow{\text{swimming}} + \overrightarrow{\text{cycling}} + \overrightarrow{\text{running}} \approx \overrightarrow{\text{triathlon}}$ is present in our embedding, despite the fact that we didn't train any ternary relationships at all. Finally, we note that there are interesting patterns in embedding quality in the hyperparameter space — even small changes in parameters can lead to sharp drops in embedding quality, as can be seen in the ridge that cuts across the middle of Figure 2. This suggests it is important to understand where in the hyperparameter space one's embedding is.

Generating community embeddings helps us overcome two major obstacles to measuring activity diversity. First, we can now measure the similarity between any pair of communities by calculating the cosine similarity between their vector representations.

[1]We have made our sets of ground-truth community analogies available at https://osf.io/jcz3p/.

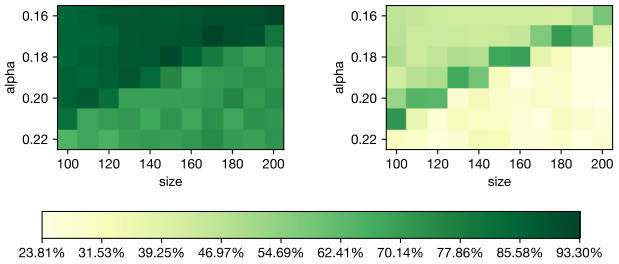


Figure 2: Heatmap of Precision@1 analogy test scores as a function of embedding algorithm hyperparameters on university to city (left) and North American sports (right).

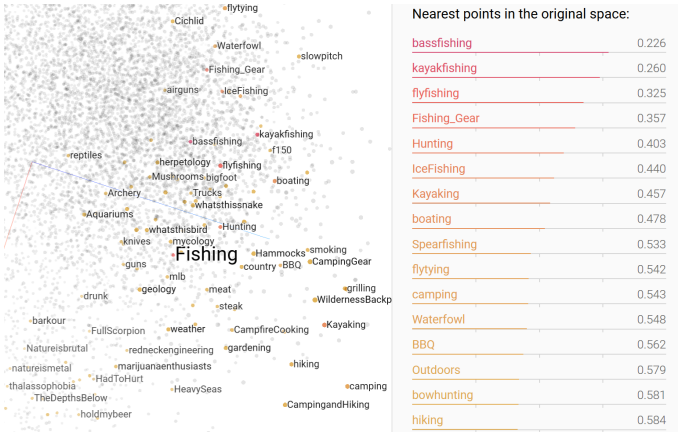


Figure 3: Two-dimensional projection of our final embedding, centered on `r/Fishing`, after performing hyperparameter search with our community analogies.

This similarity is entirely data-driven, and is in a way crowdsourced: two communities are similar if and only if many people choose to comment in them both. Other methods, such as hand-crafted taxonomies or expert classifications, are potentially biased by subjective opinions. Second, our similarities are consistent with each other, since they are derived from cosine similarities between vectors embedded in the same vector space.

2.3 Measuring Activity Diversity

We are now ready to define our measure of activity diversity. Consider a user who is active in some subset of communities. Now, if we consider their activity through the lens of our community embedding, their activity is distributed through the space according to where their communities are. Intuitively, a specialist user's activity will be concentrated in this space, as their communities are similar to each other, and a generalist user's activity will be diffuse in this space, as their communities are different from each other.

With this intuition in mind, we formulate the *Generalist-Specialist score*, or *GS-score*, as follows. Say user u_i makes w_j contributions to community c_j, and in a slight abuse of notation let $\vec{c_j}$ denote c_j's vector representation in our community embedding (note we

normalize all vectors to be unit length). First we define u_i's *center of mass* to be $\vec{\mu_i} = \sum_j w_j \vec{c_j}$. Then the definition of u_i's activity diversity is very natural — it is simply the average cosine similarity between u_i's communities and his center of mass, weighted by number of contributions by community:

$$GS(u_i) = \frac{1}{J} \sum_j w_j \frac{\vec{c_j} \cdot \vec{\mu_i}}{||\mu_i||}$$

where J is the number of communities u_i is active in. When the context is clear we will refer to u_i's GS-score as GS_i. A simple interpretation of the GS-score is that it is equal to the expected cosine similarity between a user's contributions and his center of mass in the community embedding (where we say a contribution is located at the vector representation of the community it was contributed to). The GS-score ranges between -1 and 1, where -1 is an extreme generalist and 1 is an extreme specialist. In practice, the GS-score typically ranges between 0.5 and 1.

This definition fulfills a number of desiderata. First, as the communities one contributes to become more similar to each other, the measure increases and moves towards the specialist end of the spectrum. Second, as users contribute more heavily in one community, they become more specialist. Third, fixing a user's contributions and adding another contribution to a new community pushes the measure towards the generalist end.

Interestingly, the GS-score is related to another natural definition of activity diversity. Instead of comparing contributions to the center of mass, we could imagine comparing communities to each other. Consider the *all-pairs community similarity*: $\frac{1}{C} \sum_i \sum_j \vec{c_i} \cdot \vec{c_j}$, where here the sums are over contributions, $\vec{c_i}$ is the vector representing the community of the i-th contribution, and C is the total number of contributions. This quantity measures the expected similarity between two of a user's contributions. This definition is also intuitive: the closer a user's contributions are on average, the most specialist the user. After simple algebra, we derive that the all-pairs community similarity is equivalent to the square of the GS-score. This is encouraging: two different ways of defining activity diversity are rank-equivalent, which indicates that we are capturing a real property of a user's distribution over activities.

Validating the GS-score. The GS-score is natural, intuitive, and is rank-equivalent with another intuitive definition. How well does it capture human judgments of generalists and specialists? To answer this question, we developed an online quiz asking people to compare two Reddit users at a time and decide which one they think is more specialist. By eliciting human judgments in this way, we can compare human responses against our measure and measure the extent to which the two agree.

In order to elicit high-quality feedback, we developed the quiz as a game, rather than paying for responses on a crowdsourcing platform, under the assumption that people would be more inclined to answer in good faith if they were internally motivated than if they were externally motivated. When a user arrived at the quiz, they were shown an initial screen explaining, "In this quiz, you will be presented with a list of communities (subreddits) someone is a part of. The bar shows what percentage of their activity is in this community. At each step, you will be shown two users. Click the button below the user which is **more of a specialist**. A specialist

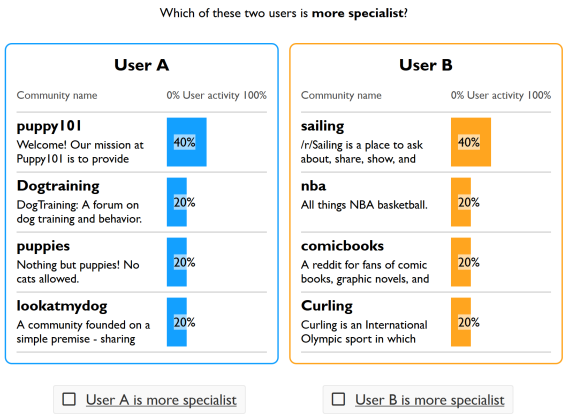


Figure 4: Screenshot of the quiz asking people to rank which user is more specialist.

is **someone whose interests focus on a particular topic**." On every subsequent screen, they were presented with one question, for example the one shown in Figure 4. Every question compares two users with exactly 4 communities each, shows the name and description of each user's communities, and the distribution of their contributions among these communities. The quiz-taker is then asked to select which user they think is more specialist.

In all, we elicited judgments on 1,807 pairs of users from 144 unique quiz-takers. Examining the relationship between how these judgments compare with the GS-score, we observe a strong correlation between the difference in GS-scores of pairs of users and the proportion of quiz-takers who agree with the GS-score (i.e. say the user with the higher GS-score is more specialist). As the difference in GS-scores grows, human judgments become more in line with the score. This is exactly what we want: as two users' activity diversity scores become closer together, it is more of a close call, and thus human judgments are less unanimous. As they become further apart, human judgments are more unanimous, and approach 90% agreement with the GS-score. Recall that even when the GS-scores of a pair of users are similar, the underlying communities they are contributing to are often very different. Thus it is reassuring that when we predict their activity diversity to be quite close, humans have a correspondingly difficult time telling them apart. The Pearson correlation between the difference in GS-scores and proportion of quiz-takers agreeing with the GS-score is 0.94. The GS-score thus accords very well with human judgments of activity diversity.

3 ACTIVITY DIVERSITY IN ONLINE PLATFORMS

Now that we have defined and validated a measure of activity diversity in online platforms, we apply our metric and study user behavior through this lens.

3.1 User Activity Diversity

Distribution of GS-scores. First, we calculate the GS-scores of all users in our Reddit and GitHub datasets. Since we're interested in the activity diversity of those who are users of the platform itself, as

opposed to "drive-by" users who may only be aware of individual communities and not the entire platform, we restrict to users who have contributed to at least three communities in their lifetime[2]. The distributions of users' GS-scores broken down by number of communities are shown in Figure 5 (top).

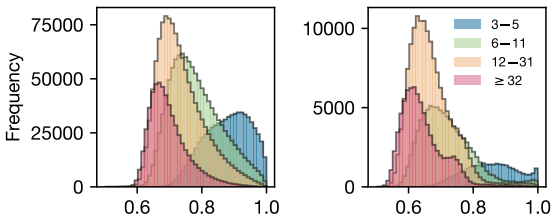
First and foremost, we observe a broad diversity of user types, ranging from extreme generalists to extreme specialists, on both Reddit and GitHub. For example, a Reddit user with GS = 0.96 contributed to communities `wedding`, `weddingplanning`, `JustEngaged`, whereas another user with GS = 0.63 contributed to communities `zelda`, `australia`, `bisexual`, `horror`, `AskMen`, `homeland`. We also see that the GS-score is related to the number of communities one participates in, as desired. Generally speaking, the more communities one contributes to, the more generalist one is. However, this is not always the case, also as desired. In fact, we observe a broad diversity of GS-scores even when we fix the number of communities. All of the distributions in Reddit, for example, have support that is almost as broad support as all of them together — even if you contribute to 30 communities, say, you can still be an extreme generalist or an extreme specialist, depending on how similar the communities you contribute to are. We see many examples where users with 32+ communities are more specialist than users with 5 communities.

To check that these results are meaningful, we compare against a null hypothesis. We conduct the following permutation test: for each user, we keep their number of communities the same, but draw the communities independently from the overall distribution of communities weighted by popularity. Are the empirical results we observe simply produced by the distributions of community popularity and number of communities by user?

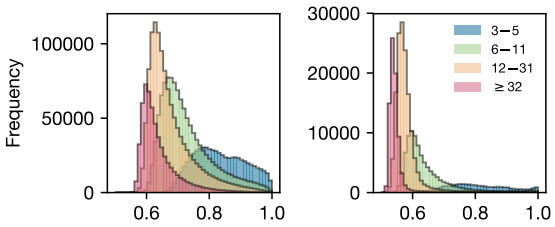
In fact, they are not. In Figure 5 (bottom), we plot the distributions of GS-scores produced by this null hypothesis. Comparing the empirical distributions with the null hypothesis distributions, we observe several major differences. We see that there are significantly more specialists at every number of communities, especially on GitHub, in the empirical data than in the permuted data. We also observe much more overlap between the GS-scores of users with different numbers of communities in the real data. Finally, we see that the most generalist constructed users are only marginally more generalist than the most generalist real users, indicating that real users are approaching the mathematical limit of how generalist they can be given the number of communities they are in.

Relationship with entropy. In previous work, the most popular method of measuring activity diversity is using the entropy of a user's activity distribution. We examine the relationship between the GS-score and entropy. As desired, they are related: the more specialist users have lower entropy. But there is still significant overlap in the entropy distributions of different GS-scores; the GS-score can distinguish between users with similar activity distributions but over sets of communities with different levels of coherence. The R^2 between GS-score and entropy is 0.69 in both Reddit and GitHub, meaning that a significant amount of the variance in GS-score (31%) is unexplained by entropy.

[2]One could use the GS-score to study the activity diversity of those who only contributed to one or two communities, but it isn't the focus of our work here.



(a) Empirical distributions of the GS-score.



(b) Null hypothesis distributions of the GS-score.

Figure 5: Empirical (top) and null hypothesis (bottom) distributions of GS-scores over users, broken down by number of communities, in Reddit (left) and GitHub (right).

Feedback success. On Reddit, every comment can be upvoted or downvoted by other users, indicating the audience's reaction to it. We use this as a proxy for success and ask how success relates with activity diversity. The raw distribution of success as a function of user GS-score is very flat, indicating that the range of viable strategies on Reddit is broad. However, there are many potential confounding factors influencing a comment's vote score: the post it is in reply to, how long after the original post it was written, its parent comment's score, the audience that sees it, etc. Thus, we consider the following question: does a comment get a higher score than its parent? (As comments on Reddit are threaded, they have some parent they are replying to, except for top-level comments). This elegantly controls for several factors. Parent and child comments usually occur close together in time, are in direct conversation with each other and so are more likely to share topics, and if the child outscores the parent (the parent usually scores higher due to its increased exposure) it can be considered a success. We examine the probability that a comment outscores its parent as a function of the author's GS-score in Figure 6. There is a significant, moderate effect, with specialists being 20% more likely than generalists to author a comment that outscores its parent. This is consistent with specialists being more engaged with, and more successful in, their home communities.

Longevity in communities. How long do users stay engaged in a community that they've joined? Is there a difference between how likely specialists and generalists are to stay? To answer this question, we examine all the occurrences of when Reddit and GitHub users join a new community, and calculate the likelihood that they subsequently make contributions in that community for six months in a row. Of course, this probability will be low on an absolute scale, but we are interested in whether generalist and specialists

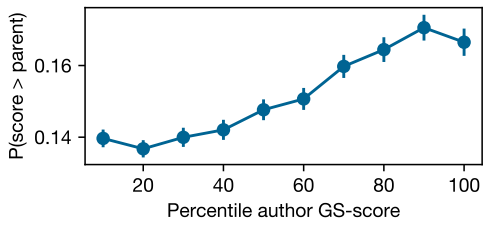


Figure 6: Comment success versus author GS-score.

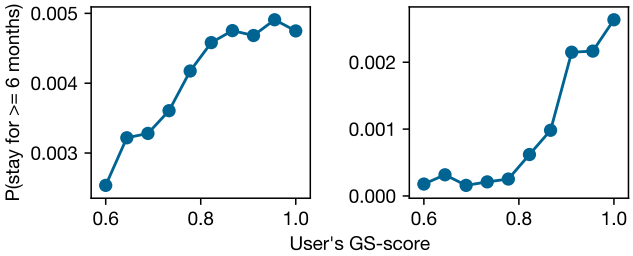


Figure 7: Likelihood of being engaged in a community for six consecutive months after first contributing, as a function of user GS-score on Reddit (left) and GitHub (right).

act differently. We plot this relationship in Figure 7. There is a clear, monotonically increasing relationship between the user's GS-score and the likelihood that they remain engaged for at least six months. The difference is quite large; on Reddit, specialists are about twice as likely to stay as generalists, and on GitHub the difference is even larger, with specialists about 10 times more likely than generalists to remain engaged. This reveals one way in which generalists and specialists are systematically different: generalists explore, specialists exploit. Generalists are more likely to try a new community but then leave, whereas specialists are more likely to stay in communities that they join. This has implications for community managers: given that a specialists have a much higher chance of converting into long-time members, it may be worth expending more effort or resources in recruiting specialists.

Longevity on the platform. We have just seen that specialists are more likely to stay in communities they join. What about the entire platform itself? Since activity is such a major factor in whether people will stay (as it is a sign that they are engaged), we control for activity level. Concretely, we consider users' activity in the first half of 2015, and measure their GS-scores on this activity only. We then measure the probability that they are active two years later, in the second half of 2017, as a function of GS-score and activity.

These results are shown in Figure 8, in which there are a number of notable findings. On Reddit, there is a very large difference between generalists and specialists in the probability of leaving the platform. For any given level of activity, specialists (e.g. 4th quartile of GS-score, shown in red) are much less likely to remain on the platform than generalists (e.g. 1st quartile of GS-score, shown in blue). For example, among users with 20 contributions during the first half of 2015, generalists remain on Reddit 75% of the time while specialists remain only 55% of the time. Even more strikingly, small differences in the GS-score can trump large differences in activity.

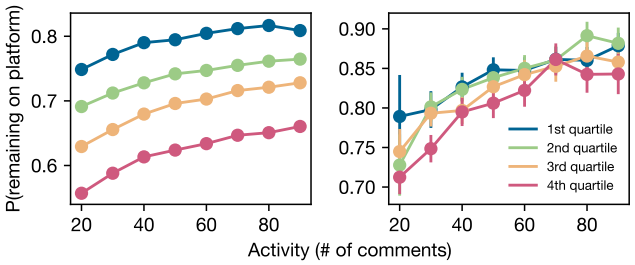


Figure 8: Likelihood of remaining on the platform versus user GS-score on Reddit (left) and GitHub (right).

For example, users with 4th-quartile GS-scores (specialists) and 90 comments only remain 66% of the time, whereas users with 2nd-quartile GS-scores and only 20 comments remain 69% of the time. Finally, we note that the probability of remaining is monotonically increasing in activity for every quartile of GS-score. On GitHub, the differences between generalists and specialists are considerably less stark. However, specialists are still significantly less likely than generalists to remain on the platform, given activity level.

At a high level, these two groups of results suggest there is an interesting interplay between activity diversity and longevity. While specialists are more likely to stay in a given community once they've tried it, generalists are more likely to stay on the entire platform. These results have important implications for designers of online platforms. Community managers may prefer exposure to specialists to maximize the chance of acquiring new long-term members. But all else equal, generalists are more likely to be long-standing members of the platform itself. Thus, it may be in the platform's interest to expose users to a more diverse set of communities. Also, designing online platforms to attract generalists and be conducive to community exploration could be beneficial.

Diversity of exposure. Every user in a platform engages with some subset of the platform's population. Depending on the communities one contributes to, the subpopulation one is exposed to can vary dramatically, and can be quite different from the platform population as a whole. How does the diversity of the group one is exposed to vary with activity diversity? We conduct an analysis on Reddit, where for each user i we consider the set of users P_i that they replied to (comments on Reddit are threaded, and thus are often in reply to each other). We call this population of users one interacts with the "parent-universe". A key benefit of our definition of activity diversity is that it can be applied to more than a user. Recall that the definition is an average over contributions, and so far we have been considering the set of contributions made by a user. Here we consider all contributions by users in the set P_i, and measure the activity diversity of this whole set. This is the diversity of the subpopulation user i is exposed to.

In Figure 9, we show how distributions of $GS(P_i)$ vary as a function of $GS(u_i)$ — how the diversity of the subpopulation one interacts with changes a function of one's own activity diversity. We observe a clear, strong trend: generalists are exposed to a diverse set of users, whereas specialists are exposed to more homogeneous sets of users. Users with first-decile GS-scores are exposed to a subpopulation of users with combined $GS(P_i)$ around 0.61, whereas users with tenth-decile GS-scores are exposed to a subpopulation of

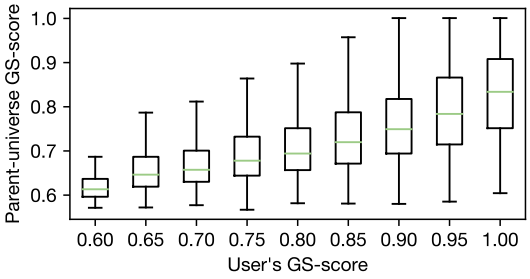


Figure 9: Distribution of the parent-universe GS-scores as a function of user GS-score.

users with combined $GS(P_i)$ around 0.84, a huge difference. Specialists, by selecting small, coherent sets of communities to contribute to, limit themselves to relatively narrow "echo chambers", where the users they interact with are themselves very specialist in their collective interests. While this trend is intuitive, the magnitude of the effect is striking.

In this paper, we are focusing on the entire set of a user's contributions, but there are many interesting analyses to be done by concentrating on a topical subset. For example, one could straightforwardly apply the GS-score to only user contributions to overtly political subreddits. Repeating this analysis would then measure the extent to which generalists and specialists are living in political echo chambers online. We leave this and other analyses as examples of promising future work using our methodology and score.

3.2 Community Activity Diversity

We measure a user's activity diversity by considering how similar their communities are. Once we understand the activity diversity of every user in a community, we can use this to also understand the activity diversity of a *community*. Here, we define the GS-score $GS(c_i)$ of a community c_i as the weighted average over its users: $GS(c_i) = \frac{1}{N} \sum_j w_j \cdot GS(u_j)$, where u_j makes w_j contributions to c_i and there are $N = \sum_j w_j$ contributions in total. Specialist (generalist) communities are those with specialist (generalist) users on average. What kinds of communities tend to be generalist or specialist? Examining the distributions over community GS-scores, we observe a broad range of community activity diversity on both Reddit and GitHub, with most communities skewing generalist.

Community GS-score stability. Does the community GS-score capture a real property of a community? One way to test this is to measure if it is at least stable. In Figure 10, we plot how community GS-scores change over time. We separate the communities into quartiles based on their initial GS-score, then plot how the distribution of their GS-scores evolves over a 3-year period on Reddit and a 1-year period on GitHub (GitHub communities are more volatile and don't last as long on average). In both Reddit and GitHub, they are strikingly consistent. The quartiles are essentially non-overlapping in their community GS-scores for the entire period. We see strong evidence that generalist communities remain generalist and specialist communities remain specialist. Furthermore, we see evidence that communities have stable GS-scores even when the underlying user population changes dramatically. For example, the subreddit `r/gardening` has highly cyclical usage patterns over the

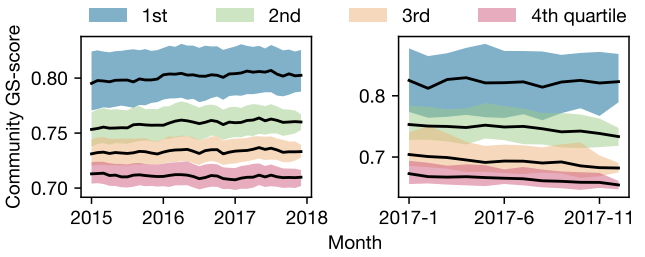


Figure 10: Stability of community GS-scores over time on Reddit (left) and GitHub (right).

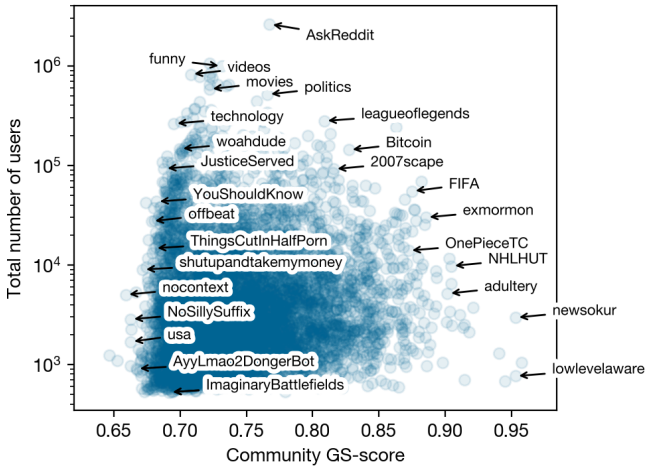


Figure 11: Community size versus community GS-score.

year, being much more popular during the summer. Despite the volume of active users changing by an order of magnitude throughout the year, the community GS-score is very consistent.

Size vs. community GS-score. Finally, we observe how community size (number of users) and community GS-score relate. In Figure 11, we show a scatterplot of this relationship on Reddit (GitHub is similar). First, we observe a broad range of community GS-scores for almost every level of community size. Even communities with tens of thousands of users can have vastly different GS-scores. For example, on Reddit, `r/The_Donald` and `r/books` have exactly the same number of users during 2017 (241K), but `r/books` has a GS-score of 0.72 and `r/The_Donald` has a GS-score of 0.86, which are on opposite ends of the community activity diversity spectrum. Inspecting the communities, the GS-scores are intuitive: generalist communities are typically about very broad topics like `r/technology` or have wide appeal like `r/funny`, whereas specialist communities are typically centered around particular interests or niches like `r/FIFA` and `r/exmormon`. On GitHub, generalist communities are general-use frameworks, such as `d3` (a web-based data visualization framework), whereas specialist communities are more deeply technical, such as `grpc` (a remote procedure call system).

Elite users vs. community GS-score. Finally, we combine the user GS-score and community GS-score. Who are the top users in generalist and specialist communities? In each community, we rank

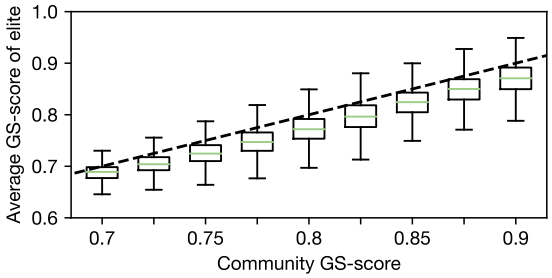


Figure 12: Elite users' GS-scores versus community GS-score.

users by the average comment vote score, and consider the top 5% as the "elites". Are the generalists or specialists the elites, and how does this vary with community GS-score? The relationship, shown in Figure 12, indicates that generalists are consistently more likely to be among the elites. The line $y = x$ indicates what would happen under a random baseline, and clearly the distributions are always below this line. We speculate that this is due to generalists having more chances at being in the elite of a community. By our measure, a user can be in a community's elite with a small number of highly-rated comments. Since generalists are more likely to spread their activity out over more communities, they have more opportunities to be in the elite.

4 PREDICTION TASKS

In the previous Sections, we introduced our methodology for quantifying activity diversity and applied it to study user behavior in online platforms. Now we demonstrate that our insights are directly applicable to two important prediction problems in online platforms: community recommendation and user retention. We perform these tasks on our main dataset, Reddit.

4.1 Community Recommendation

Predicting which communities a user is likely to join next is a ubiquitous and important prediction task for online platforms. Our first task is to predict which new communities users will join after an initial observation period. On this task, every model outputs a ranking of communities for each user sorted by the likelihood of that user joining. We trained the models on all data from 2017 and tested them on the communities users joined for the first time during January 2018. We evaluated four models:

- **Random**: a baseline that randomly ranks communities.
- **Popularity**: communities ranked by popularity
- **Collaborative Filtering**: a state-of-the-art model (Bayesian Personalized Ranking [20]) for doing collaborative filtering with implicit data.
- **Center-of-Mass Nearest Neighbors**: a novel method using community embedding, which simply ranks communities by how cosine-similar they are to a user's center of mass during the training period (2017). Note that a community embedding trained on only 2017 data was used to ensure that the CF model and our embedding-based model had access to the exact same training data.

To assess model performance, we looked at users who were active in the training period and the test period. On Reddit, there were

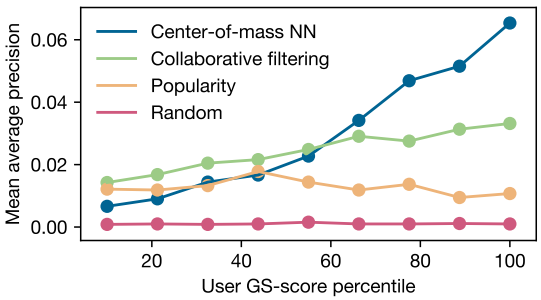


Figure 13: Performance of community recommendation models on Reddit versus user GS-score percentile.

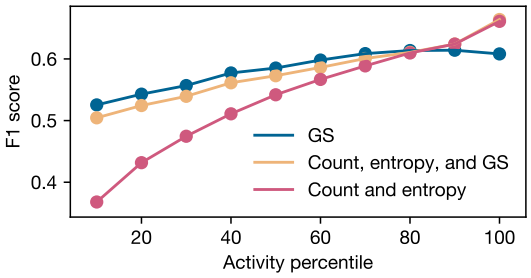


Figure 14: Performance of the Reddit longevity prediction models on users of varying activity levels.

20K such users with 24K total community joins. We use mean average precision (MAP) to assess the quality of community rankings produced by each model. The overall results are: 0.001 for random, 0.013 for popularity, 0.023 for CF, and 0.027 for Center-of-Mass NN. Despite being an extremely simple model, Center-of-Mass NN slightly outperforms a state-of-the-art collaborative filtering method, which demonstrates the fidelity of our community embeddings. Notice that our method doesn't even take popularity into account (which CF does).

Beyond this proof of concept, we are interested in how predictability varies with user GS-score. This relationship is shown in Figure 13, and there are two notable findings. First, it is clear that Center-of-Mass NN is competitive with CF among all users, but especially outperforms it on specialists. Second, it is remarkable that both CF and Center-of-Mass NN perform much better for specialist users. The fact that some users are so much more predictable than others, and that this correlates so well with the GS-score, is interesting and has important implications. For example, the value of recommendations could be much higher for specialists, which should inform their cost.

4.2 Platform User Retention

Predicting whether users will remain or leave the platform is another ubiquitous and important problem. Here, we demonstrate that the GS-score is valuable in solving this problem. Concretely, the task is to consider all users in an initial period and predict whether or not they will be present on the platform in a later period. We used the first half of 2015 as the initial period and the second half of 2017 as the later period, and randomly sample a balanced set

of 464K users to consider (so that 50% stay and 50% leave). We split this into a training set and test set of equal size, and measure our models with the F1 score. We train a simple logistic regression model with comment count, entropy, and user GS-score as features. We compare against entropy since it is the main metric others have used in past work to quantify activity diversity.

We see that the GS-score alone outperforms the other models. Over most of the range of activity, but particularly for low-activity users, the improvement over entropy is quite large: for users in the lowest decile of activity, the F1 score of activity and entropy is 0.36 but the GS-score alone achieves an F1 score of 0.53 (see Figure 14). Count and entropy outperform the GS-score for the highest decile since activity is such a strong predictor of engagement.

5 RELATED WORK

Our work draws on three areas of previous research: embedding models, analyses of user behavior in social platforms, and work investigating generalists and specialists more broadly.

Embedding models. Our work draws upon the embedding methods that have precipitated revolutions in many fields, starting with natural language processing. Word embedding models like word2vec showed how to use modern model architectures to leverage large datasets and create embeddings that preserve semantic relationships [16, 17]. Subsequent work by Levy and Goldberg improved our understanding of these models, and also provided software that we used to generate our community embeddings [13, 14]. Kumar et al. applied word embedding algorithms to develop community embeddings in their study of community conflict on Reddit [11], and early steps in this direction were taken by Martin [15]. Embedding methods have also been developed for use in the context of social networks [9, 19, 24].

User behavior in social platforms. Our work is part of a large body of work investigating user behavior in social platforms. Reddit in particular has been the subject of a rich line of research. Tan and Lee study the trajectory of communities that Reddit users post in and find that users settle after an initial period of exploration [23]. They (and others) use entropy as a measure of topical concentration; as mentioned in Section 2, entropy is agnostic to the similarity or dissimilarity of different activities. Our measure and analysis builds on this work by taking community similarity into account. Hamilton et al. study community loyalty, both from the user and community perspectives [10], and Zhang et al. study community identity on Reddit, finding differences between communities with high and low churn rates [26].

Generalists and specialists. The notion of a spectrum spanning from generalists to specialists is ancient. It begins with the epigraph that opens this paper—"The fox knows many things; but the hedgehog knows one big thing."—written by 7th-century BC Greek poet Archilocus. This quote was the inspiration for the classic essay *The Hedgehog and the Fox* by philosopher Isaiah Berlin, in which he classifies writers and thinkers as either hedgehogs, those who have a single unifying idea, or foxes, those who draw on a wide array of experiences [4]. His work, in turn, brought the idea of generalists and specialists back into mainstream thought, and it has served as a useful frame ever since. As one example, Tetlock, in his studies of expert judgment and forecasters, analyzed the predictive abilities of generalists and specialists ("hedgehogs" and "foxes"), finding that specialists performed worse, especially on long-term forecasts within their specialty [25]. The problem of how broadly to allocate one's energy is ubiquitous: in biology, species are broadly categorized into generalist and specialist species [8]. Generalist species can thrive in a wide variety of environments, whereas specialist species are limited to a narrow range of environments.

In the study of activity diversity in online platforms, the closest work is Adamic et al.'s research on the relationship between individual focus and contribution on Wikipedia [1]. Our work builds on theirs by introducing a higher-fidelity metric for activity similarity based on community embeddings, studying other online platforms, assigning activity diversity scores to communities, and investigating the relationship between activity diversity and additional outcome metrics. Anderson studied how skill diversity correlates with worker pay in an online crowdwork platform [2].

6 CONCLUSION

In this work, we have introduced a methodology to quantify activity diversity in online platforms, and analyzed several important phenomena around user behavior through this lens on Reddit and GitHub. We developed high-quality *community embeddings*, which encode similarities between communities as defined by their members, as well as a set of community analogies to validate them. Based on these embeddings, we introduced the *GS-score*, a principled measure of activity diversity. We found a broad spectrum of user styles, from extreme generalists to extreme specialists, and observed that specialists are more likely to produce higher-quality replies. We also observed that specialists are much more likely to stay in communities they contribute to, but generalists are much more likely to remain on the platform as a whole. We applied activity diversity to study an important question: how diverse are the subpopulations users engage with? We found that generalists engage with a significantly more diverse group of people, whereas specialists are exposed to much narrower segments of the population. We also studied the activity diversity of communities, and found that the community GS-score is remarkably stable over time, even when the underlying user base changes a lot. In our prediction tasks, we found that specialists are more predictable than generalists in which communities they will join, and that the user GS-score is much more predictive of user retention than entropy.

There are several promising directions for future work. First, the GS-score can be directly applied to answer important questions about *subsets* of user activity in online platforms. For example, it would be interesting to conduct our results on echo chambers in the subset of political activity on Reddit. The GS-score then becomes a measure of political diversity. Second, as our GS-score metric is robust and conceptually simple, it could be broadly applicable beyond the online platforms we considered in this work. It would be illuminating to apply the GS-score to other domains as well.

Acknowledgments. This work was partially supported by NSERC. We thank Josef Waller, Shunzhe Yu, and especially Will Hamilton for fruitful discussions, and we thank Jason Baumgartner and the GH Archive for providing public access to the Reddit and GitHub datasets, respectively.

REFERENCES

[1] Lada A Adamic, Xiao Wei, Jiang Yang, Sean Gerrish, Kevin K Nam, and Gavin S Clarkson. 2010. Individual focus and knowledge contribution. *arXiv preprint arXiv:1002.0561* (2010).

[2] Katharine A Anderson. 2017. Skill networks and measures of complex human capital. *Proceedings of the National Academy of Sciences (PNAS)* (2017).

[3] Jason Baumgartner. 2017. pushshift.io Reddit archive. https://pushshift.io/. (2017). Accessed: 2018-07-23.

[4] Isaiah Berlin. 1953. *The hedgehog and the fox.* Weidenfeld & Nicolson.

[5] Hendrik Bode, Frederick Mosteller, John W Tukey, and Charles Winsor. 1949. The education of a scientific generalist. *Science* (1949).

[6] Elisabeth Bublitz and Florian Noseleit. 2014. The skill balancing act: when does broad expertise pay off? *Small Business Economics* 42, 1 (2014).

[7] Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).*

[8] Nicholas B Davies, John R Krebs, and Stuart A West. 2012. *An introduction to behavioural ecology.* John Wiley & Sons.

[9] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NIPS).*

[10] William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM).*

[11] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference (WWW).*

[12] Erin Leahey. 2007. Not by productivity alone: How visibility and specialization contribute to academic earnings. *American Sociological Review* 72, 4 (2007), 533–561.

[13] Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).*

[14] Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems (NIPS).*

[15] Trevor Martin. 2017. community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the Second Workshop on NLP and Computational Social Science.*

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS).*

[18] Scott Page. 2007. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies.* Princeton University Press.

[19] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).*

[20] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI).*

[21] Kurt C Stange. 2009. The generalist approach. *The Annals of Family Medicine* 7, 3 (2009).

[22] Rosemary Stevens. 2017. *Medical practice in modern England: the impact of specialization and state medicine.* Routledge.

[23] Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web (WWW).*

[24] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web (WWW).*

[25] Philip E Tetlock. 2005. *Expert political judgment: How good is it? How can we know?* Princeton University Press.

[26] Justine Zhang, William L Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM).*