Community Identity and User Engagement in a Multi-Community Landscape

Justine Zhang*
Cornell University
jz727@cornell.edu

William L. Hamilton*
Stanford University
wleif@stanford.edu

Cristian Danescu-Niculescu-Mizil
Cornell University
cristian@cs.cornell.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Abstract

A community's identity defines and shapes its internal dynamics. Our current understanding of this interplay is mostly limited to glimpses gathered from isolated studies of individual communities. In this work we provide a systematic exploration of the nature of this relation across a wide variety of online communities. To this end we introduce a quantitative, language-based typology reflecting two key aspects of a community's identity: how *distinctive*, and how temporally *dynamic* it is. By mapping almost 300 Reddit communities into the landscape induced by this typology, we reveal regularities in how patterns of user engagement vary with the characteristics of a community.

Our results suggest that the way new and existing users engage with a community depends strongly and systematically on the nature of the collective identity it fosters, in ways that are highly consequential to community maintainers. For example, communities with distinctive and highly dynamic identities are more likely to retain their users. However, such niche communities also exhibit much larger acculturation gaps between existing users and newcomers, which potentially hinder the integration of the latter.

More generally, our methodology reveals differences in how various social phenomena manifest across communities, and shows that structuring the multi-community landscape can lead to a better understanding of the systematic nature of this diversity.

1 Introduction

"If each city is like a game of chess, the day when I have learned the rules, I shall finally possess my empire, even if I shall never succeed in knowing all the cities it contains."

— Italo Calvino, Invisible Cities

A community's identity—defined through the common interests and shared experiences of its users—shapes various facets of the social dynamics within it (Ren, Kraut, and Kiesler 2007; Tajfel 2010; Ren et al. 2012). Numerous instances of this interplay between a community's identity and social dynamics have been extensively studied in the context of individual online communities (Bryant, Forte, and Bruckman 2005; Lampe et al. 2010; Danescu-Niculescu-Mizil et al. 2013). However, the sheer variety of online platforms complicates the task of generalizing insights beyond these isolated, single-community glimpses. A new way to reason about the variation across multiple communities is needed in order to systematically characterize the relationship between properties of a community and the dynamics taking place within.

One especially important component of community dynamics is user engagement. We can aim to understand why users join certain communities (Panciera, Halfaker, and Terveen 2009), what factors influence user retention (Dror et al. 2012), and how users react to innovation (Danescu-Niculescu-Mizil et al. 2013). While striking patterns of user engagement have been uncovered in prior case studies of individual communities (Postmes, Spears, and Lea 2000; Huffaker et al. 2006; Fugelstad et al. 2012; Otterbacher and Hemphill 2012; McAuley and Leskovec 2013), we do not know whether these observations hold beyond these cases, or when we can draw analogies between different communities. Are there certain types of communities where we can expect similar or contrasting engagement patterns?

To address such questions quantitatively we need to provide structure to the diverse and complex space of online communities. Organizing the multi-community landscape would allow us to both characterize individual points within this space, and reason about systematic variations in patterns of user engagement across the space.

Present work: Structuring the multi-community space. In order to systematically understand the relationship between community identity[1] and user engagement we introduce a quantitative typology of online communities. Our typology is based on two key aspects of community identity: how *distinctive*—or niche—a community's interests are relative to other communities, and how *dynamic*—or volatile—these interests are over time. These axes aim to capture the salience of a community's identity and dynamics of its temporal evolution.

*The two first authors contributed equally and are ordered non-alphabetically to balance ordering in another collaboration.

[1] We use "community identity" and "collective identity" interchangeably to refer to the shared definition of a group, derived from members' common interests and shared experiences. We are not directly concerned with the more sociopolitical and psychological connotations of these terms (Polletta and Jasper 2001; Simon and Klandermans 2001; Ashmore, Deaux, and McLaughlin-Volpe 2004).

Our main insight in implementing this typology automatically and at scale is that the *language* used within a community can simultaneously capture how distinctive and dynamic its interests are. This language-based approach draws on a wealth of literature characterizing linguistic variation in online communities and its relationship to community and user identity (Cassell and Tversky 2005; Danescu-Niculescu-Mizil et al. 2013; Bamman, Eisenstein, and Schnoebelen 2014; Tran and Ostendorf 2016; Eisenstein 2017). Basing our typology on language is also convenient since it renders our framework immediately applicable to a wide variety of online communities, where communication is primarily recorded in a textual format.

Using our framework, we map almost 300 Reddit communities onto the landscape defined by the two axes of our typology (Section 2). We find that this mapping induces conceptually sound categorizations that effectively capture key aspects of community-level social dynamics. In particular, we quantitatively validate the effectiveness of our mapping by showing that our two-dimensional typology encodes signals that are predictive of community-level rates of user retention, complementing strong activity-based features.

Engagement and community identity. We apply our framework to understand how two important aspects of user engagement in a community—the community's propensity to retain its users (Section 3), and its permeability to new members (Section 4)—vary according to the type of collective identity it fosters. We find that communities that are characterized by specialized, constantly-updating content have higher user retention rates, but also exhibit larger linguistic gaps that separate newcomers from established members.

More closely examining factors that could contribute to this linguistic gap, we find that especially within distinctive communities, established users have an increased propensity to engage with the community's specialized content, compared to newcomers (Section 5). Interestingly, while established members of distinctive communities more avidly respond to temporal updates than newcomers, in more generic communities it is the *outsiders* who engage more with volatile content, perhaps suggesting that such content may serve as an entry-point to the community (but not necessarily a reason to stay). Such insights into the relation between collective identity and user engagement can be informative to community maintainers seeking to better understand growth patterns within their online communities.

More generally, our methodology stands as an example of how sociological questions can be addressed in a multi-community setting. In performing our analyses across a rich variety of communities, we reveal both the diversity of phenomena that can occur, as well as the systematic nature of this diversity.

2 A typology of community identity

A community's identity derives from its members' common interests and shared experiences (Ashmore, Deaux, and McLaughlin-Volpe 2004; Ritzer 2007). In this work, we structure the multi-community landscape along these two key dimensions of community identity: how *distinctive* a community's interests are, and how *dynamic* the community is over time.

We now proceed to outline our quantitative typology, which maps communities along these two dimensions. We start by providing an intuition through inspecting a few example communities. We then introduce a generalizable language-based methodology and use it to map a large set of Reddit communities onto the landscape defined by our typology of community identity.

2.1 Overview and intuition

In order to illustrate the diversity within the multi-community space, and to provide an intuition for the underlying structure captured by the proposed typology, we first examine a few example communities and draw attention to some key social dynamics that occur within them.

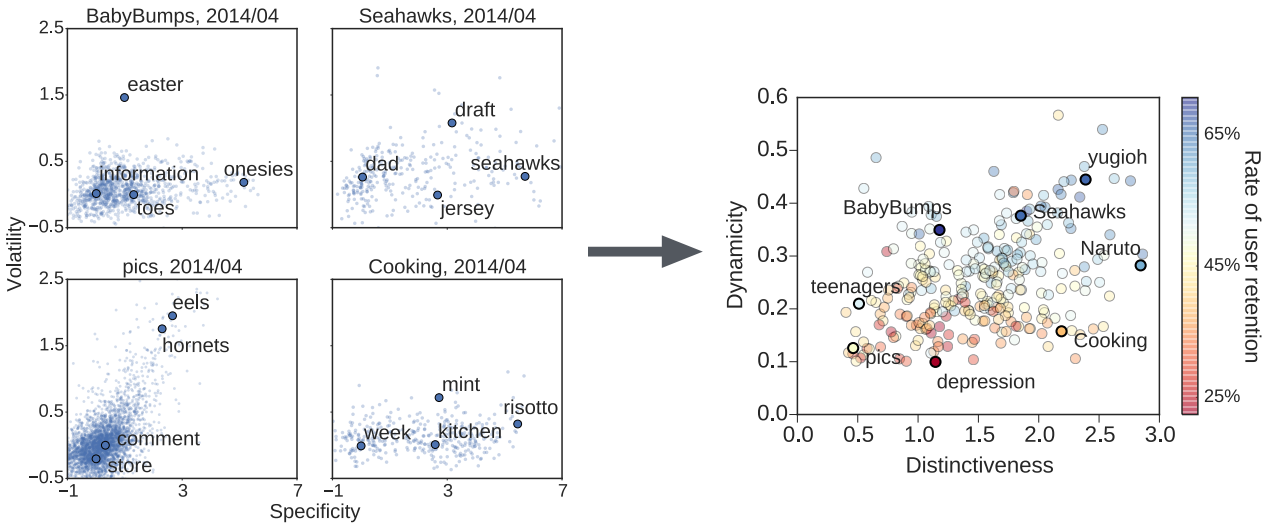
We consider four communities from Reddit: in *Seahawks*, fans of the Seahawks football team gather to discuss games and players; in *BabyBumps*, expecting mothers trade advice and updates on their pregnancy; *Cooking* consists of recipe ideas and general discussion about cooking; while in *pics*, users share various images of random things (like eels and hornets). We note that these communities are topically contrasting and foster fairly disjoint user bases. Additionally, these communities exhibit varied patterns of user engagement. While Seahawks maintains a devoted set of users from month to month, pics is dominated by transient users who post a few times and then depart.

Discussions *within* these communities also span varied sets of interests. Some of these interests are more specific to the community than others: *risotto*, for example, is seldom a discussion point beyond Cooking. Additionally, some interests consistently recur, while others are specific to a particular time: *kitchen*s are a consistent focus point for cooking, but *mint* is only in season during spring. Coupling specificity and consistency we find interests such as *easter*, which isn't particularly specific to BabyBumps but gains prominence in that community around Easter (see Figure 1.A for further examples).

These specific interests provide a window into the nature of the communities' interests as a whole, and by extension their community identities. Overall, discussions in Cooking focus on topics which are highly distinctive and consistently recur (like *risotto*). In contrast, discussions in Seahawks are highly dynamic, rapidly shifting over time as new games occur and players are traded in and out. In the remainder of this section we formally introduce a methodology for mapping communities in this space defined by their *distinctiveness* and *dynamicity* (examples in Figure 1.B).

2.2 Language-based formalization

Our approach follows the intuition that a distinctive community will use language that is particularly *specific*, or unique, to that community. Similarly, a dynamic community will use *volatile* language that rapidly changes across successive windows of time. To capture this intuition automatically, we start by defining word-level measures of specificity and volatility. We then extend these word-level primitives to characterize entire comments, and the community itself.



A. Within-community, word-level measures **B. Community-level measures**

Figure 1: **A**: Within a community certain words are more community-specific and temporally volatile than others. For instance, words like *onesies* are highly specific to the *BabyBumps* community (top left), while words like *easter* are temporally ephemeral. **B**: Extending these word-level measures to communities, we can measure the overall distinctiveness and dynamicity of a community, which are highly associated with user retention rates (colored heatmap; see Section 3). Communities like *Seahawks* (a football team) and *Cooking* use highly distinctive language. Moreover, *Seahawks* uses very dynamic language, as the discussion continually shifts throughout the football season. In contrast, the content of *Cooking* remains stable over time, as does the content of *pics*; though these communities do have ephemeral fads, the overall themes discussed generally remain consistent.

Our characterizations of words in a community are motivated by methodology from prior literature that compares the frequency of a word in a particular setting to its frequency in some background distribution, in order to identify instances of linguistic variation (Monroe, Colaresi, and Quinn 2008; Eisenstein 2017). Our particular framework makes this comparison by way of pointwise mutual information (PMI).

In the following, we use c to denote one community within a set $\mathcal{C}$ of communities, and t to denote one time period within the entire history T of $\mathcal{C}$. We account for temporal as well as inter-community variation by computing word-level measures for each time period of each community's history, c_t. Given a word w used within a particular community c at time t, we define two word-level measures:

Specificity. We quantify the *specificity* $\mathcal{S}_c(w)$ of w to c by calculating the PMI of w and c, relative to $\mathcal{C}$,

$$\mathcal{S}_c(w) = \log \frac{\mathrm{P}_c(w)}{\mathrm{P}_{\mathcal{C}}(w)},$$

where $\mathrm{P}_c(w)$ is w's frequency in c. w is *specific* to c if it occurs more frequently in c than in the entire set $\mathcal{C}$, hence distinguishing this community from the rest. A word w whose occurrence is decoupled from c, and thus has $\mathcal{S}_c(w)$ close to 0, is said to be *generic*.

We compute values of $\mathcal{S}_{c_t}(w)$ for each time period t in T; in the above description we drop the time-based subscripts for clarity.

Volatility. We quantify the *volatility* $\mathcal{V}_{c_t}(w)$ of w to c_t as the PMI of w and c_t relative to c_T, the entire history of c:

$$\mathcal{V}_{c_t}(w) = \log \frac{\mathrm{P}_{c_t}(w)}{\mathrm{P}_{c_T}(w)}.$$

A word w is *volatile* at time t in c if it occurs more frequently at t than in the entire history T, behaving as a fad within a small window of time. A word that occurs with similar frequency across time, and hence has $\mathcal{V}$ close to 0, is said to be *stable*.

Extending to utterances. Using our word-level primitives, we define the specificity of an utterance d in c, $\mathcal{S}_c(d)$ as the average specificity of each word in the utterance. The volatility of utterances is defined analogously.

2.3 Community-level measures

Having described these word-level measures, we now proceed to establish the primary axes of our typology:

Distinctiveness. A community with a very distinctive identity will tend to have distinctive interests, expressed through specialized language. Formally, we define the distinctiveness of a community $\mathcal{N}(c_t)$ as the average specificity of all utterances in c_t. We refer to a community with a less distinctive identity as being *generic*.

Dynamicity. A highly dynamic community constantly shifts interests from one time window to another, and these temporal variations are reflected in its use of volatile language.

	generic	distinctive
dynamic	BabyBumps IAmA Libertarian australia	CollegeBasketball Seahawks formula1 yugioh
consistent	AdviceAnimals funny news pics	Cooking Guitar MakeupAddiction harrypotter

Table 1: Examples of communities on Reddit which occur at the extremes (top and bottom quartiles) of our typology.

Formally, we define the dynamicity of a community $\mathcal{D}(c_t)$ as the average volatility of all utterances in c_t. We refer to a community whose language is relatively consistent throughout time as being *stable*.

In our subsequent analyses, we focus mostly on examing the *average* distinctiveness and dynamicity of a community over time, denoted $\mathcal{N}(c)$ and $\mathcal{D}(c)$.

2.4 Applying the typology to Reddit

We now explain how our typology can be applied to the particular setting of Reddit, and describe the overall behaviour of our linguistic axes in this context.

Dataset description. Reddit is a popular website where users form and participate in discussion-based communities called *subreddits*. Within these communities, users post content—such as images, URLs, or questions—which often spark vibrant lengthy discussions in thread-based comment sections.

The website contains many highly active subreddits with thousands of active subscribers. These communities span an extremely rich variety of topical interests, as represented by the examples described earlier. They also vary along a rich multitude of structural dimensions, such as the number of users, the amount of conversation and social interaction, and the social norms determining which types of content become popular. The diversity and scope of Reddit's multi-community ecosystem make it an ideal landscape in which to closely examine the relation between varying community identities and social dynamics.

Our full dataset consists of all subreddits on Reddit from January 2013 to December 2014,[2] for which there are at least 500 words in the vocabulary used to estimate our measures, in at least 4 months of the subreddit's history. We compute our measures over the comments written by users in a community in time windows of *months*, for each sufficiently active month, and manually remove communities where the bulk of the contributions are in a foreign language. This results in 283 communities (c), for a total of 4,872 community-months (c_t).[3]

[2] https://archive.org/details/2015_reddit_comments_corpus

[3] While we chose these cutoffs on the dataset to ensure robust estimates of the linguistic measures, we note that slight relaxations produce qualitatively similar results in the later analyses.

Estimating linguistic measures. We estimate word frequencies $\mathrm{P}_{c_t}(w)$, and by extension each downstream measure, in a carefully controlled manner in order to ensure we capture robust and meaningful linguistic behaviour. First, we only consider top-level comments which are initial responses to a post, as the content of lower-level responses might reflect conventions of dialogue more than a community's high-level interests. Next, in order to prevent a few highly active users from dominating our frequency estimates, we count each unique word *once* per user, ignoring successive uses of the same word by the same user. This ensures that our word-level characterizations are not skewed by a small subset of highly active contributors.[4]

In our subsequent analyses, we will only look at these measures computed over the *nouns* used in comments. In principle, our framework can be applied to any choice of vocabulary. However, in the case of Reddit using nouns provides a convenient degree of interpretability. We can easily understand the implication of a community preferentially mentioning a noun such as *gamer* or *feminist*, but interpreting the overuse of verbs or function words such as *take* or *of* is less straightforward. Additionally, in focusing on nouns we adopt the view emphasized in modern "third wave" accounts of sociolinguistic variation, that stylistic variation is inseparable from topical content (Eckert 2012). In the case of online communities, the choice of what people choose to talk about serves as a primary signal of social identity. That said, a typology based on more purely stylistic differences is an interesting avenue for future work.

Accounting for rare words. One complication when using measures such as PMI, which are based off of ratios of frequencies, is that estimates for very infrequent words could be overemphasized (Turney and Littman 2003). Words that only appear a few times in a community tend to score at the extreme ends of our measures (e.g. as highly specific or highly generic), obfuscating the impact of more frequent words in the community. To address this issue, we discard the long tail of infrequent words in our analyses, using only the top 5th percentile of words, by frequency within each c_t, to score comments and communities.[5]

Typology output on Reddit. The distribution of $\mathcal{N}$ and $\mathcal{D}$ across Reddit communities is shown in Figure 1.B, along with examples of communities at the extremes of our typology. We find that interpretable groupings of communities emerge at various points within our axes. For instance, highly distinctive and dynamic communities tend to focus on rapidly-updating interests like sports teams and games, while generic and consistent communities tend to be large "link-sharing" hubs where users generally post content with no clear dominating themes. More examples of communities at the extremes of our typology are shown in Table 1.

[4] Understanding the role that highly active users (Hamilton et al. 2017) play in shaping a community's dynamics is an interesting direction for future work.

[5] For the purposes of the present analyses, this method produces reasonable output that is robust to small variations in our choice of parameters. However, it would be fruitful in future work to consider other methods, e.g., (Monroe, Colaresi, and Quinn 2008), for capturing linguistic variation.

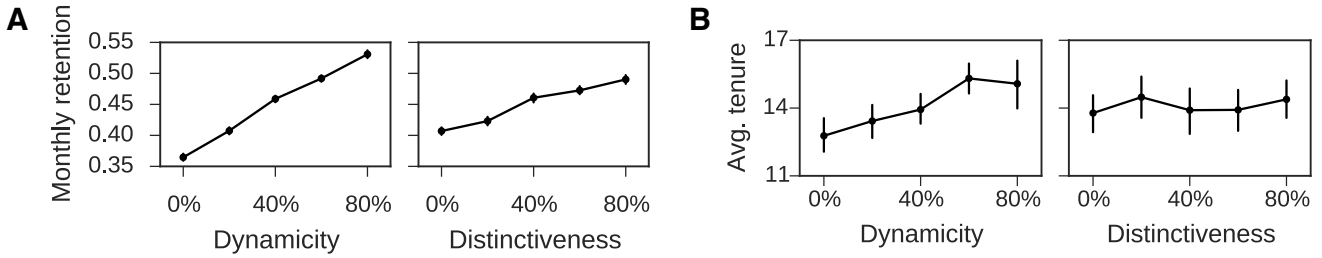


Figure 2: **A**: The monthly retention rate for communities differs drastically according to their position in our identity-based typology, with dynamicity being the strongest signal of higher user retention (x-axes bin community-months by percentiles; in all subsequent plots, error bars indicate 95% bootstrapped confidence intervals). **B**: Dynamicity also correlates with long-term user retention, measured as the number of months the average user spends in the community; however, distinctiveness does not correlate with this longer-term variant of user retention.

We note that these groupings capture abstract properties of a community's content that go beyond its topic. For instance, our typology relates topically contrasting communities such as *yugioh* (which is about a popular trading card game) and *Seahawks* through the shared trait that their content is particularly distinctive. Additionally, the axes can clarify differences between topically similar communities: while *startrek* and *thewalkingdead* both focus on TV shows, *startrek* is less dynamic than the median community, while *thewalkingdead* is among the most dynamic communities, as the show was still airing during the years considered.

3 Community identity and user retention

We have seen that our typology produces qualitatively satisfying groupings of communities according to the nature of their collective identity. This section shows that there is an informative and highly predictive relationship between a community's position in this typology and its user engagement patterns. We find that communities with distinctive and dynamic identities have higher rates of user engagement, and further show that a community's position in our identity-based landscape holds important predictive information that is complementary to a strong activity baseline.

In particular user retention is one of the most crucial aspects of engagement and is critical to community maintenance (Ren et al. 2012). We quantify how successful communities are at retaining users in terms of both short and long-term commitment. Our results indicate that rates of user retention vary drastically, yet systematically according to how distinctive and dynamic a community is (Figure 1).

We find a strong, explanatory relationship between the temporal consistency of a community's identity and rates of user engagement: dynamic communities that continually update and renew their discussion content tend to have far higher rates of user engagement. The relationship between distinctiveness and engagement is less universal, but still highly informative: niche communities tend to engender strong, focused interest from users at one particular point in time, though this does not necessarily translate into long-term retention.

3.1 Community-type and monthly retention

We find that dynamic communities, such as *Seahawks* or *starcraft*, have substantially higher rates of monthly user retention than more stable communities (Spearman's ρ = 0.70, p <0.001, computed with community points averaged over months; Figure 2.A, left). Similarly, more distinctive communities, like *Cooking* and *Naruto*, exhibit moderately higher monthly retention rates than more generic communities (Spearman's ρ = 0.33, p <0.001; Figure 2.A, right).

Monthly retention is formally defined as the proportion of users who contribute in month t and then return to contribute again in month $t + 1$. Each monthly datapoint is treated as unique and the trends in Figure 2 show 95% bootstrapped confidence intervals, cluster-resampled at the level of subreddit (Field and Welsh 2007), to account for differences in the number of months each subreddit contributes to the data.

Importantly, we find that in the task of predicting community-level user retention our identity-based typology holds additional predictive value on top of strong baseline features based on community-size (# contributing users) and activity levels (mean # contributions per user), which are commonly used for churn prediction (Dror et al. 2012). We compared out-of-sample predictive performance via leave-one-community-out cross validation using random forest regressors with ensembles of size 100, and otherwise default hyperparameters (Pedregosa et al. 2011). A model predicting average monthly retention based on a community's average distinctiveness and dynamicity achieves an average mean squared error (MSE) of 0.0060 and $R^2 = 0.37$,[6] while an analogous model predicting based on a community's size and average activity level (both log-transformed) achieves MSE $= 0.0062$ and $R^2 = 0.35$. The difference between the two models is not statistically significant ($p = 0.99$, Wilcoxon signed-rank test). However, combining features from both models results in a large and statistically significant improvement over each independent model (MSE $= 0.0038$, $R^2 = 0.60$, $p < 0.001$ Bonferroni-corrected pairwise Wilcoxon tests). These results indicate that our typology can explain variance in community-level retention rates, and provides information beyond what is present in standard activity-based features.

[6]We measure out-of-sample R^2 relative to a baseline that predicts the mean of the training data (Campbell and Thompson 2008).

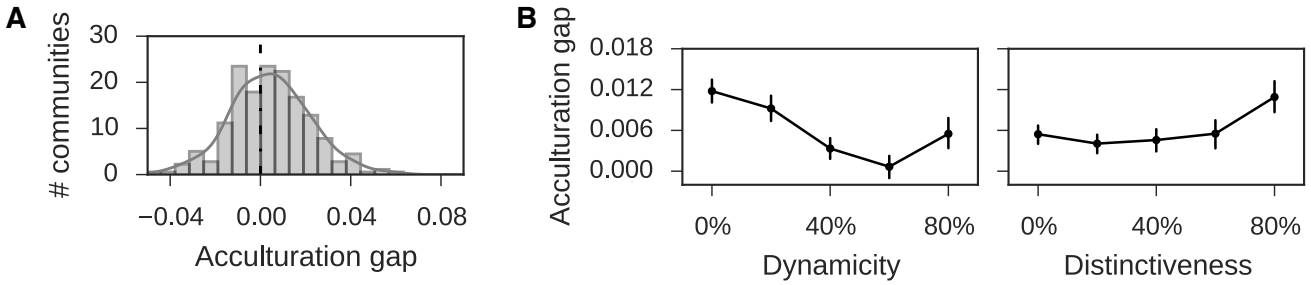


Figure 3: **A**: There is substantial variation in the direction and magnitude of the acculturation gap, which quantifies the extent to which established members of a community are linguistically differentiated from outsiders. Among 60% of communities this gap is positive, indicating that established users match the community's language more than outsiders. **B**: The size of the acculturation gap varies systematically according to how dynamic and distinctive a community is. Distinctive communities exhibit larger gaps; as do relatively stable, and very dynamic communities.

3.2 Community-type and user tenure

As with monthly retention, we find a strong positive relationship between a community's dynamicity and the average number of months that a user will stay in that community (Spearman's $\rho = 0.41$, $p < 0.001$, computed over all community points; Figure 2.B, left). This verifies that the short-term trend observed for monthly retention translates into longer-term engagement and suggests that long-term user retention might be strongly driven by the extent to which a community continually provides novel content. Interestingly, there is no significant relationship between distinctiveness and long-term engagement (Spearman's $\rho = 0.03$, $p = 0.77$; Figure 2.B, right). Thus, while highly distinctive communities like *RandomActsOfMakeup* may generate focused commitment from users over a short period of time, such communities are unlikely to retain long-term users unless they also have sufficiently dynamic content.

To measure user tenures we focused on one slice of data (May, 2013) and measured how many months a user spends in each community, on average—the average number of months between a user's first and last comment in each community.[7] We have activity data up until May 2015, so the maximum tenure is 24 months in this set-up, which is exceptionally long relative to the average community member (throughout our entire data less than 3% of users have tenures of more than 24 months in any community).

4 Community identity and acculturation

The previous section shows that there is a strong connection between the nature of a community's identity and its basic user engagement patterns. In this section, we probe the relationship between a community's identity and how permeable, or accessible, it is to outsiders.

We measure this phenomenon using what we call the *acculturation gap*, which compares the extent to which engaged vs. non-engaged users employ community-specific language. While previous work has found this gap to be large and predictive of future user engagement in two beer-review communities (Danescu-Niculescu-Mizil et al. 2013), we find that the size of the acculturation gap depends strongly on the nature of a community's identity, with the gap being most pronounced in stable, highly distinctive communities (Figure 3).

This finding has important implications for our understanding of online communities. Though many works have analyzed the dynamics of "linguistic belonging" in online communities (Cassell and Tversky 2005; Nguyen and Rose 2011; Danescu-Niculescu-Mizil et al. 2013; Bamman, Eisenstein, and Schnoebelen 2014), our results suggest that the process of linguistically fitting in is highly contingent on the nature of a community's identity. At one extreme, in generic communities like *pics* or *worldnews* there is no distinctive, linguistic identity for users to adopt.

To measure the acculturation gap for a community, we follow Danescu-Niculescu-Mizil et al (2013) and build "snapshot language models" (SLMs) for each community, which capture the linguistic state of a community at one point of time.[8] Using these language models we can capture how linguistically close a particular utterance is to the community by measuring the cross-entropy of this utterance relative to the SLM:

$$H(d, \text{SLM}_{c_t}) = \frac{1}{|d|} \sum_{b_i \in d} \text{SLM}_{c_t}(b_i), \qquad (1)$$

where $\text{SLM}_{c_t}(b_i)$ is the probability assigned to bigram b_i from comment d in community-month c_t. We build the SLMs by randomly sampling 200 active users—defined as users with at least 5 comments in the respective community and month. For each of these 200 active users we select 5 random 10-word spans from 5 unique comments.[9] To ensure robustness and maximize data efficiency, we construct 100 SLMs for each community-month pair that has enough data, bootstrap-resampling from the set of active users.

[7] Analogous results hold for other reasonable choices of month.

[8] We use Katz-Backoff bigram language models with Good-Turing smoothing (Chen and Goodman 1999) and vocabularies of size 50,000.

[9] Using fixed-length spans controls for spurious length-effects (Danescu-Niculescu-Mizil et al. 2013); the same controls are used in the cross-entropy calculations.

We compute a basic measure of the acculturation gap for a community-month c_t as the relative difference of the cross-entropy of comments by users active in c_t with that of singleton comments by *outsiders*—i.e., users who only ever commented once in c, but who are still active[10] in Reddit in general:

$$A(c_t) = \frac{\mathbb{E}_{d\sim\mathcal{V}_s}[H(d, \text{SLM}_{c_t})] - \mathbb{E}_{d\sim\mathcal{V}_a}[H(d, \text{SLM}_{c_t})]}{\mathbb{E}_{d\sim\mathcal{V}_a}[H(d, \text{SLM}_{c_t})]}. \quad (2)$$

$\mathcal{V}_s$ denotes the distribution over singleton comments, $\mathcal{V}_a$ denotes the distribution over comments from users active in c_t, and $\mathbb{E}$ the expected values of the cross-entropy over these respective distributions. For each bootstrap-sampled SLM we compute the cross-entropy of 50 comments by active users (10 comments from 5 randomly sampled active users, who were not used to construct the SLM) and 50 comments from randomly-sampled outsiders.

Figure 3.A shows that the acculturation gap varies substantially with how distinctive and dynamic a community is. Highly distinctive communities have far higher acculturation gaps, while dynamicity exhibits a non-linear relationship: relatively stable communities have a higher linguistic 'entry barrier', as do very dynamic ones. Thus, in communities like *IAmA* (a general Q&A forum) that are very generic, with content that is highly, but not extremely dynamic, outsiders are at no disadvantage in matching the community's language. In contrast, the acculturation gap is large in stable, distinctive communities like *Cooking* that have consistent community-specific language. The gap is also large in *extremely* dynamic communities like *Seahawks*, which perhaps require more attention or interest on the part of active users to keep up-to-date with recent trends in content.

These results show that phenomena like the acculturation gap, which were previously observed in individual communities (Nguyen and Rose 2011; Danescu-Niculescu-Mizil et al. 2013), cannot be easily generalized to a larger, heterogeneous set of communities. At the same time, we see that structuring the space of possible communities enables us to observe systematic patterns in how such phenomena vary.

5 Community identity and content affinity

Through the acculturation gap, we have shown that communities exhibit large yet systematic variations in their permeability to outsiders. We now turn to understanding the divide in commenting behaviour between outsiders and active community members at a finer granularity, by focusing on two particular ways in which such gaps might manifest among users: through different levels of engagement with *specific* content and with temporally *volatile* content.

Echoing previous results, we find that community type mediates the extent and nature of the divide in content affinity. While in distinctive communities active members have a higher affinity for both community-specific content and for highly volatile content, the opposite is true for generic communities, where it is the outsiders who engage more with volatile content.

[10]Users must comment at least 5 times in a month to be considered active in Reddit.

We quantify these divides in content affinity by measuring differences in the language of the comments written by active users and outsiders. Concretely, for each community c, we define the *specificity gap* $\Delta\mathcal{S}_c$ as the relative difference between the average specificity of comments authored by active members, and by outsiders, where these measures are macroaveraged over users. Large, positive $\Delta\mathcal{S}_c$ then occur in communities where active users tend to engage with substantially more community-specific content than outsiders.

We analogously define the *volatility gap* $\Delta\mathcal{V}_c$ as the relative difference in volatilities of active member and outsider comments. Large, positive values of $\Delta\mathcal{V}_c$ characterize communities where active users tend to have more volatile interests than outsiders, while negative values indicate communities where active users tend to have more stable interests.

We find that in 94% of communities, $\Delta\mathcal{S}_c > 0$, indicating (somewhat unsurprisingly) that in almost all communities, active users tend to engage with more community-specific content than outsiders. However, the magnitude of this divide can vary greatly: for instance, in *Homebrewing*, which is dedicated to brewing beer, the divide is very pronounced ($\Delta\mathcal{S}_c = 0.33$) compared to *funny*, a large hub where users share humorous content ($\Delta\mathcal{S}_c = 0.011$).

The nature of the volatility gap is comparatively more varied. In *Homebrewing* ($\Delta\mathcal{V}_c = 0.16$), as in 68% of communities, active users tend to write more volatile comments than outsiders ($\Delta\mathcal{V}_c > 0$). However, communities like *funny* ($\Delta\mathcal{V}_c$ = -0.16), where active users contribute relatively stable comments compared to outsiders ($\Delta\mathcal{V}_c < 0$), are also well-represented on Reddit.

To understand whether these variations manifest systematically across communities, we examine the relationship between divides in content affinity and community type. In particular, following the intuition that active users have a relatively high affinity for a community's niche, we expect that the distinctiveness of a community will be a salient mediator of specificity and volatility gaps. Indeed, we find a strong correlation between a community's distinctiveness and its specificity gap (Spearman's $\rho = 0.34$, $p < 0.001$).

We also find a strong correlation between distinctiveness and community volatility gaps (Spearman's $\rho = 0.53$, $p < 0.001$). In particular, we see that among the most *distinctive* communities (i.e., the top third of communities by distinctiveness), active users tend to write more volatile comments than outsiders (mean $\Delta\mathcal{V}_c = 0.098$), while across the most *generic* communities (i.e., the bottom third), active users tend to write more *stable* comments (mean $\Delta\mathcal{V}_c$ = -0.047, Mann-Whitney U test $p < 0.001$). The relative affinity of outsiders for volatile content in these communities indicates that temporally ephemeral content might serve as an entry point into such a community, without necessarily engaging users in the long term.

6 Further related work

Our language-based typology and analysis of user engagement draws on and contributes to several distinct research threads, in addition to the many foundational studies cited in the previous sections.

Multicommunity studies. Our investigation of user engagement in multicommunity settings follows prior literature which has examined differences in user and community dynamics across various online groups, such as email listservs. Such studies have primarily related variations in user behaviour to structural features such as group size and volume of content (Butler 2001; Jones, Ravid, and Rafaeli 2004; Backstrom et al. 2008; Kairam, Wang, and Leskovec 2012). In focusing on the linguistic content of communities, we extend this research by providing a *content-based* framework through which user engagement can be examined.

Reddit has been a particularly useful setting for studying multiple communities in prior work. Such studies have mostly focused on characterizing how individual *users* engage across a multi-community platform (Tan and Lee 2015; Hessel, Tan, and Lee 2016), or on specific user engagement patterns such as loyalty to particular communities (Hamilton et al. 2017). We complement these studies by seeking to understand how features of *communities* can mediate a broad array of user engagement patterns within them.

Typologies of online communities. Prior attempts to typologize online communities have primarily been qualitative and based on hand-designed categories, making them difficult to apply at scale. These typologies often hinge on having some well-defined function the community serves, such as supporting a business or non-profit cause (Porter 2004), which can be difficult or impossible to identify in massive, anonymous multi-community settings. Other typologies emphasize differences in communication platforms and other functional requirements (Preece 2001; Stanoevska-Slabeva and Schmid 2001), which are important but preclude analyzing differences between communities within the same multi-community platform. Similarly, previous computational methods of characterizing multiple communities have relied on the presence of markers such as affixes in community names (Hessel, Tan, and Lee 2016), or platform-specific affordances such as evaluation mechanisms (Lee, Jin, and Mimno 2016).

Our typology is also distinguished from community detection techniques that rely on structural or functional categorizations (Leskovec et al. 2008; Yang and Leskovec 2015). While the focus of those studies is to identify and characterize sub-communities within a larger social network, our typology provides a characterization of predefined communities based on the nature of their identity.

Broader work on collective identity. Our focus on community identity dovetails with a long line of research on collective identity and user engagement, in both online and offline communities (Allen and Meyer 1996; Tajfel 2010; Ren et al. 2012). These studies focus on individual-level psychological manifestations of collective (or social) identity, and their relationship to user engagement (Allen and Meyer 1996; Meyer et al. 2002; Utz 2003; Ren, Kraut, and Kiesler 2007).

In contrast, we seek to characterize community identities at an aggregate level and in an interpretable manner, with the goal of systematically organizing the diverse space of online communities. Typologies of this kind are critical to these broader, social-psychological studies of collective identity: they allow researchers to systematically analyze how the psychological manifestations and implications of collective identity vary across diverse sets of communities.

7 Conclusion and future work

Our current understanding of engagement patterns in online communities is patched up from glimpses offered by several disparate studies focusing on a few individual communities. This work calls into attention the need for a method to systematically reason about similarities and differences across communities. By proposing a way to structure the multi-community space, we find not only that radically contrasting engagement patterns emerge in different parts of this space, but also that this variation can be at least partly explained by the type of identity each community fosters.

Our choice in this work is to structure the multi-community space according to a typology based on community identity, as reflected in language use. We show that this effectively explains cross-community variation of three different user engagement measures—retention, acculturation and content affinity—and complements measures based on activity and size with additional interpretable information. For example, we find that in niche communities established members are more likely to engage with volatile content than outsiders, while the opposite is true in generic communities. Such insights can be useful for community maintainers seeking to understand engagement patterns in their own communities.

One main area of future research is to examine the *temporal dynamics* in the multi-community landscape. By averaging our measures of distinctiveness and dynamicity across time, our present study treated community identity as a static property. However, as communities experience internal changes and respond to external events, we can expect the nature of their identity to shift as well. For instance, the relative consistency of *harrypotter* may be disrupted by the release of a new novel, while *Seahawks* may foster different identities during and between football seasons. Conversely, a community's type may also mediate the impact of new events. Moving beyond a static view of community identity could enable us to better understand how temporal phenomena such as linguistic change manifest across different communities, and also provide a more nuanced view of user engagement—for instance, are communities more welcoming to newcomers at certain points in their lifecycle?

Another important avenue of future work is to explore other ways of mapping the landscape of online communities. For example, combining structural properties of communities (Leskovec et al. 2008) with topical information (Hessel, Tan, and Lee 2016) and with our identity-based measures could further characterize and explain variations in user engagement patterns. Furthermore, extending the present analyses to even more diverse communities supported by different platforms (e.g., GitHub, StackExchange, Wikipedia) could enable the characterization of more complex behavioral patterns such as collaboration and altruism, which become salient in different multicommunity landscapes.

Acknowledgements

The authors thank Liye Fu, Jack Hessel, David Jurgens and Lillian Lee for their helpful comments. This research has been supported in part by a Discovery and Innovation Research Seed Award from the Office of the Vice Provost for Research at Cornell, NSF CNS-1010921, IIS-1149837, IIS-1514268 NIH BD2K, ARO MURI, DARPA XDATA, DARPA SIMPLEX, DARPA NGS2, Stanford Data Science Initiative, SAP Stanford Graduate Fellowship, NSERC PGS-D, Boeing, Lightspeed, and Volkswagen.

References

Allen, N. J., and Meyer, J. P. 1996. Affective, continuance, and normative commitment to the organization: An examination of construct validity. *J. Vocational Behavior*.

Ashmore, R. D.; Deaux, K.; and McLaughlin-Volpe, T. 2004. An organizing framework for collective identity: Articulation and significance of multidimensionality. *Psych. Bulletin*.

Backstrom, L.; Kumar, R.; Marlow, C.; Novak, J.; and Tomkins, A. 2008. Preferential behavior in online groups. In *Proceedings of WSDM*.

Bamman, D.; Eisenstein, J.; and Schnoebelen, T. 2014. Gender identity and lexical variation in social media. *J. Sociolinguistics*.

Bryant, S. L.; Forte, A.; and Bruckman, A. 2005. Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proceedings of GROUP*.

Butler, B. S. 2001. Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Info. Sys. Research*.

Campbell, J. Y., and Thompson, S. B. 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*.

Cassell, J., and Tversky, D. 2005. The language of online intercultural community formation. *J. CMC*.

Chen, S., and Goodman, J. 1999. An empirical study of smoothing techniques for language modeling. *Comp. Speech & Lang*.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of WWW*.

Dror, G.; Pelleg, D.; Rokhlenko, O.; and Szpektor, I. 2012. Churn prediction in new users of Yahoo! answers. In *Proceedings of WWW*.

Eckert, P. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annu. Rev. Anthro.*

Eisenstein, J. 2017. Written dialect variation in online social media. In Boberg, C.; Nerbonne, J.; and Watt, D., eds., *Handbook of Dialectology*. Wiley.

Field, C. A., and Welsh, A. H. 2007. Bootstrapping clustered data. *J. Royal Stat. Soc.: Series B (Statistical Methodology)*.

Fugelstad, P.; Dwyer, P.; Filson Moses, J.; Kim, J.; Mannino, C. A.; Terveen, L.; and Snyder, M. 2012. What makes users rate (share, tag, edit...)?: Predicting patterns of participation in online communities. In *Proceedings of CSCW*.

Hamilton, W.; Zhang, J.; Danescu-Niculescu-Mizil, C.; Jurafsky, D.; and Leskovec, J. 2017. Loyalty in online communities. In *Proceedings of ICWSM*.

Hessel, J.; Tan, C.; and Lee, L. 2016. Science, AskScience, and BadScience: On the coexistence of highly related communities. In *Proceedings of ICWSM*.

Huffaker, D.; Jorgensen, J.; Iacobelli, F.; Tepper, P.; and Cassell, J. 2006. Computational measures for language similarity across time in online communities. In *Proceedings of NAACL-HLT WKSP on Analyzing Conversations in Text & Speech*.

Jones, Q.; Ravid, G.; and Rafaeli, S. 2004. Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Info. Sys. Research*.

Kairam, S. R.; Wang, D. J.; and Leskovec, J. 2012. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of WSDM*.

Lampe, C.; Wash, R.; Velasquez, A.; and Ozkaya, E. 2010. Motivations to participate in online communities. In *Proceedings of CHI*.

Lee, M.; Jin, S. H.; and Mimno, D. 2016. Beyond exchangeability: the Chinese Voting Process. In *Proceedings of NIPS*.

Leskovec, J.; Lang, K. J.; Dasgupta, A.; and Mahoney, M. W. 2008. Statistical properties of community structure in large social and information networks. In *Proceedings of WWW*.

McAuley, J., and Leskovec, J. 2013. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of WWW*.

Meyer, J. P.; Stanley, D. J.; Herscovitch, L.; and Topolnytsky, L. 2002. Affective, continuance, and normative commitment to the organization: A meta-analysis of antecedents, correlates, and consequences. *J. Vocational Behavior*.

Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Pol. An.*

Nguyen, D., and Rose, C. 2011. Language use as a reflection of socialization in online communities. In *Proceedings of ACL WKSP on Languages in Soc. Media*.

Otterbacher, J., and Hemphill, L. 2012. Learning the lingo? Gender, prestige and linguistic adaptation in review communities. In *Proceedings of CSCW*.

Panciera, K.; Halfaker, A.; and Terveen, L. 2009. Wikipedians are born, not made: A study of power editors on Wikipedia. In *Proceedings of CSCW*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. scikit-learn: Machine learning in Python. *JMLR*.

Polletta, F., and Jasper, J. M. 2001. Collective identity and social movements. *Annu. Rev. Soc.*

Porter, C. E. 2004. A typology of virtual communities: A multidisciplinary foundation for future research. *J. CMC*.

Postmes, T.; Spears, R.; and Lea, M. 2000. The formation of group norms in computer-mediated communication. *Hum. Comm. Res*.

Preece, J. 2001. Sociability and usability in online communities: Determining and measuring success. *Behavior & Info. Tech.*

Ren, Y.; Kraut, R.; Kiesler, S.; and Resnick, P. 2012. Encouraging commitment in online communities. *Building successful online communities: Evidence-based social design*.

Ren, Y.; Kraut, R.; and Kiesler, S. 2007. Applying common identity and bond theory to design of online communities. *Organization studies*.

Ritzer, G. 2007. *The Blackwell Encyclopedia of Sociology*. Blackwell Publishing Malden, MA.

Simon, B., and Klandermans, B. 2001. Politicized collective identity: A social psychological analysis. *American Psychologist*.

Stanoevska-Slabeva, K., and Schmid, B. F. 2001. A typology of online communities and community supporting platforms. *System Sciences*.

Tajfel, H. 2010. *Social identity and intergroup relations*. Cambridge University Press.

Tan, C., and Lee, L. 2015. All Who Wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of WWW*.

Tran, T., and Ostendorf, M. 2016. Characterizing the language of online communities and its relation to community reception. In *Proceedings of EMNLP*.

Turney, P. D., and Littman, M. L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM TOIS*.

Utz, S. 2003. Social identification and interpersonal attraction in MUDs. *Swiss J. Psych.*

Yang, J., and Leskovec, J. 2015. Defining and evaluating network communities based on ground-truth. *Knowledge & Info. Sys.*