The long tale of the long tail: The internet as counterforce to cultural homogeneity

Abstract: The internet promised to facilitate cultural exploration, but research has found that attention allocation online resembles offline patterns in which generic communities targeting mass-market tastes attract the bulk of attention while a long tail of specialty communities targeting niche tastes remains under resourced. While past work examined how much attention different communities attracted, I instead examine the connections that form among different communities online as a function of lower structural constraints on exploration. Using data from Reddit, I independently measure the position of communities in a structural space based on membership, and a cultural space based on content. In contrast to many offline settings, I find that specialty communities online share members with even the largest, most generic communities—connections among culturally distinct communities that emerge from capturing the attention of the same users in different ways—cross-pollination that can increase exposure to niche tastes and fuel further cultural exploration.

Keywords: internet, social media, online communities, attention, culture, structure

Introduction

Research in the late 20th and early 21st centuries painted the internet as the great attention equalizer: a countervailing force against cultural homogeneity (Anderson, 2007; DiMaggio et al., 2001; Neuman, 1991). In the pre-internet world, structure—both local social networks and the structure of cultural markets—limited individuals' exploration of communities and objects for cultural expression (Fischer, 1975; Peterson and Berger 1975). In other words, structure constrained cultural exploration. In many offline, pre-internet settings, individuals settled for readily-available generic offerings and communities targeting mass-market tastes (Anderson, 2007; Brynjolfsson et al., 2011), while only a small, and often separate, group of enthusiasts put in the effort necessary to search for specialty offerings targeting niche tastes (Carroll, 1985; Carroll et al., 2002).

A key promise of the internet was to enable people to more readily engage with a long-tail of specialty offerings for cultural expression (Brynjolfsson et al., 2011). The internet lowers search costs and barriers to entry (Faraj et al., 2011), such that trying something new requires only clicking a few links or typing a few lines of text. In addition, the internet offers nearly unlimited and relatively costless space for a variety of offerings to proliferate. These differences signaled a shift from uneven distributions of attention, in which a few generic options dominate, toward a flatter distribution characterized by a thicker long-tail of specialty offerings attracting significant attention. Some research examining online marketplaces found initial empirical evidence in support of this shift (Anderson, 2007; Bhattacharjee et al., 2007; Brynjolfsson et al., 2011).

Yet other research since the beginning of the 21st century has found that the allocation of attention online resembles familiar patterns of resource allocation in offline settings. In contrast to a thicker long-tail of specialty offerings, some work found instead the reproduction of "superstar" or "rich-get-richer" effects online (Elberse and Oberholzer-Gee, 2006; Fleder and Hosanagar, 2009; Tan et al., 2017).

However, most existing work studying attention allocation online takes a distributional perspective, exploring the proportion of the total attention share that communities—or products—attract (e.g., Brynjolfsson et al., 2011). Such an approach may overlook a fundamental difference in how attention allocation occurs online: the ease of distributing and searching for content online enables individuals to spread their attention across multiple communities, in ways that would be unimaginable in the offline world (TeBlunthuis and Hill, 2022; Waller and Anderson, 2019). In other words, at the level of the individual in the online world, cultural exploration is relatively free from structural constraints of markets and local social networks. This suggests that individuals may be able to connect communities that otherwise would rely on separate members in the offline world. This matters because connections in the form of overlapping membership can generate cross-cutting discussion and encourage further exploration among individuals who may not otherwise be exposed to more specialty content in the offline world—a feedback loop of cultural exploration.

In this article, I shift from examining the distribution of attention, to instead exploring connections among communities. I ask whether different kinds of communities must rely on different users, or whether they can rely on a shared set of individuals, representing new connections enabled by capturing attention in different ways. To do so, I take an ecological approach, examining how communities targeting different tastes (i.e., generic and specialty) can survive (i.e., attract attention) based on their position in a structural space of users.

Leveraging data from the online community Reddit, a massive online platform consisting of thousands of topic-specific sub-communities known as "subreddits," I find two main results: 1) that generic and specialty communities can rely on the very same users by attracting their attention in different ways and 2) that specialty communities can find success in the most prominent locations in the ecology, connected by many overlapping users to the largest, most generic communities. Together, these findings offer evidence of diverse cultural exploration that generates connections among communities that might otherwise remain disconnected in the offline world.

Theory

Structural Constraints on Culture Offline

A key promise of the internet was to offer a more egalitarian context for social and economic life—a more level playing field for different kinds of content to emerge and find an audience, and, on the other side, for individuals to locate products and communities better matching their tastes, and thus ultimately to find ways to express their identities and preferences through consumption, social connection, and even shared production (DiMaggio et al., 2001). Stated differently, the internet promised to alter the relationship between structure—of both social networks and cultural markets—and culture—the communities and objects of consumption through which people express themselves.

Before the advent of mass communication technologies, cultural exploration was constrained by local social networks, emerging primarily through interaction resulting from physical co-location (e.g., Durkheim, 1984; Fischer, 1975). The diversity of ideas, communities, and modes of expression that any individual gained exposure to related to the diversity of that individual's local social networks. Additionally, the structure of economic markets historically constrained the diversity of cultural goods available for consumption. Research taking a production of culture perspective champions this idea (Peterson and Anand, 2004). For example, Peterson and Berger (1975) argue that when a multiplicity of firms compete for limited resources in a market, these firms must differentiate their products in order to capture some portion of the market, producing cultural diversity (Peterson and Berger, 1975; Stigler, 1952). On the other hand, when markets become concentrated, firms instead target mass-market tastes that will reach the largest number of consumers. Cultural homogeneity ensues.

Throughout the 20th century, major societal and technological shifts in the United States further tightened structure's grip on culture. Post-World War II, the proliferation of long-term, low-interest housing loans through the Federal Housing Administration (FHA), combined with new techniques for mass-producing homes and construction guidelines that favored sparsely populated areas, encouraged a massive movement of individuals from densely populated urban areas to new homes built in the suburbs (Baldassare, 1992; Nicolaides and Wiese, 2017). The interstate highway system enabled suburban sprawl to extend even further away from urban cores (Hawley, 1971). Suburbs were not only less dense than their urban predecessors, but also less diverse due to FHA underwriting guidelines requiring racial segregation until the early 1950s (Nicolaides and Wiese, 2017). This combination of sparsity and homogeneity dealt a devastating blow to the subcultural variety available to people through their local social networks (Fischer, 1975).

The structure of economic markets also shifted to further constrain culture. As discussed in literature on mass-culture theory (e.g., MacDonald, 1953; Wilensky, 1964), the explosion of mass-media in the 20th century contributed to cultural homogenization at the national level. New technologies enabled culture to travel beyond local social networks. However, increasingly concentrated markets, consisting of corporations making broad-based appeals to common-denominator tastes in an attempt to draw in newly-available national audiences, led to a "massification" of tastes (Shils, 1962). Within many cultural industries, while specialty content could survive, it was usually only available through local channels and word-of-mouth, rather than through broad marketing campaigns at the national level (Carroll, 1985; Peterson and Anand, 2004; Peterson and Berger, 1975). And so, paradoxically, the more geographic mobility people gained through advancements in transportation, and the more that culture could travel across time and space through new media technologies, the more homogenization seemed to emerge.

Structure and Culture Online

While the offline world increasingly reflected cultural homogenization and the dominance of generic cultural offerings, research in the early 21st century predicted that the emerging digital age would lead to a "long tail of the internet," specifically meaning a shift away from uneven distributions of attention and resources, in which a few generic offerings dominated, and toward a thicker, and thus more well-resourced, long tail of specialty offerings, signifying the flourishing of cultural exploration (Anderson, 2007; Brynjolfsson et al., 2011). Much of the empirical work that initially examined this shift evaluated the context of online marketplaces such as Amazon and eBay, with a number of studies finding support for the hypothesized "long tail," characterized by a more even distribution of resources between a select set of so-called "superstar" offerings and a much larger set of specialty offerings (Anderson, 2007; Brynjolfsson et al., 2011).

Several affordances of the internet fueled these predictions. On the supply side, the internet lowers costs of production, storage, and distribution—offering nearly limitless space for a variety of options to coexist, including content and communities created in a grass-roots or peer-produced fashion (Benkler and Nissenbaum, 2006). On the demand side, the internet eliminates geographic constraints, enabling individuals, including those historically disenfranchised, to find communities and content matching their preferences. The internet makes exploration easier by drastically lowering search and evaluation costs (Brynjolfsson et al., 2011). Online, individuals can evaluate fit with a community with only a few clicks of a mouse, and the fluid boundaries and low membership barriers of many online spaces facilitate rapid exploration (Faraj et al., 2011).

Despite these hopes and predictions, a large body of research paints a picture of the internet that strays from the promise of cultural exploration. In both the online marketplace contexts of many studies on the "long-tail" phenomenon, as well as in many online communities, research has found evidence of patterns of attention allocation resembling patterns of resource allocation in the offline world, stemming from the reality that, while online settings lower costs in many ways, they still rely on a limited resource—attention (Goldhaber, 1997). Namely, there is evidence that attention is allocated unevenly across offerings online, with mass-market content becoming broadly popular and attracting the bulk of attention (Elberse and Oberholzer-Gee, 2006; Johnson et al., 2014; Newman, 2003; Taeuscher, 2019; Tan et al., 2017)—reminiscent of the dominance of generic culture in many traditional offline settings (Carroll, 1985).

Explanations for these patterns revolve around the central idea that attention remains a limited resource online, and that individuals must make choices in how to allocate their attention (Lin et al., 2017; Goldhaber, 1997; Wang et al., 2013). Additionally, at the same time that an online context makes it easier for individuals to explore options, it also enables more direct access to information on peer preferences, which can bolster social influence effects leading to herding behavior that reinforces scale-free attention distributions (Adamic and Huberman, 2000)—especially in settings where recommendation algorithms funnel users toward popular content (Fleder and Hosanagar, 2009).

An Ecological Approach

Much of the work examining the allocation of attention online focuses on the amount of attention that different kinds of content attract, finding often that it is still the most culturally generic communities that attract the bulk of attention. This distributional approach may overlook an important difference in attention allocation online stemming from a looser coupling between structure and culture: specifically, the ability of individuals to spread their attention across multiple communities. This difference at the individual level has potential implications for how communities are connected by shared member resources. To investigate this possibility, I leverage an ecological lens to examine connections between online communities.

Research in organizational ecology, largely focused on traditional product markets and offline voluntary organizations, situates communities and organizations in a structural space based on the resources (e.g., consumers or members) they rely on (e.g., Hannan and Freeman, 1989; McPherson, 1983). The resource space is thus a social-structural density map of available resources, with the center representing the area of greatest resource abundance, and the periphery representing areas of relative resource scarcity (Péli and Nooteboom, 1999). Many past ecological studies of traditional markets, and even of online communities (e.g., Wang et al., 2013), implicitly assume that individual resources map onto particular cultural tastes (Carroll, 1985; Carroll et al., 2002). In other words, individuals pursue a limited set of cultural tastes and function as discrete resources which organizations compete for in a zero-sum fashion (Carroll et al., 2002).

This assumption follows from the structural constraints on cultural exploration in many offline settings. Cultural exploration in the offline world requires physical exploration of the communities and offerings available through one's local social networks. Similarly, exploration requires moving beyond the readily available, heavily marketed generic offerings catering to common-denominator tastes. In other words, pursuing different tastes is costly in the offline world (McPherson, 2004). Given this, individuals tend to restrict their activity to a cluster of related offerings, optimizing as much as they can to find the communities or "products that best match their preferences" (Péli and Nooteboom, 1999: 1133). Exploration, and sampling across tastes—omnivorousness—is a luxury restricted largely to those in the higher strata of society who possess the time, energy, and other capital necessary to do so (Peterson and Kern, 1996).

The assumption of constrained cultural exploration offline holds implications for the connections between different kinds of communities at the ecological level: communities serving generic or specialty tastes rely on primarily different resource bases, and thus occupy different positions in the overall ecology. Generic offerings (e.g., top 40 music) and the communities surrounding them exist in the center of the ecology consisting of many people holding popular tastes, and specialty offerings (e.g., death metal music) exist in the periphery consisting of a separate and smaller pool of people with niche tastes (Carroll and Hannan, 1989). Stated another way, there is a duality between a community's position in the structural resource space and its form, or the cultural taste that it serves (Breiger, 1974; McPherson, 2004; Mohr and Guerra-Pearson, 2010).

Online, there is evidence of diminished structural constraints on individual cultural exploration. Many individuals spread their attention across multiple online communities (TeBlunthuis and Hill, 2022; Waller and Anderson, 2019), given the low search costs, low barriers to entry, and fluid boundaries of online communities. Online communities can potentially capture the attention of individuals in varied, fractional ways—a breakdown in the coupling between a community's structural position and the taste it serves.

I examine whether this freedom at the individual level emerges in different kinds of connections among communities at the ecological level—specifically, whether generic and specialty

communities can capture the attention of the same users, whereas in offline settings they often must rely on different resource pools. To do so empirically, I measure the position of each community in a structural space based on shared users. I do this by measuring how much of a community's membership overlaps with other communities, and specifically the largest communities, to capture the community's proximity to the largest resource pools of membership. I also separately measure the taste targeted by the community (i.e., generic or specialty)—to account for the possibility that the taste targeted by the community is decoupled from its structural position in the user space. I then predict, based on the kind of structural position occupied and the kind of taste targeted, how communities attract attention (i.e., whether they attract a relative breadth or depth of attention). The objective is to evaluate whether both generic and specialty communities are able to coexist in highly resourced and connected areas by attracting attention of the same users in different ways—or instead, and similar to offline communities, whether only generic communities can succeed in these most prominent locations while specialty communities can only succeed in the peripheries of the ecology.

Methods

Data

Data for this study come from Reddit.com. Reddit is an online forum comprised of numerous topic specific subforums, known as subreddits, which are denoted with an "r," a forward slash, and then the subreddit title. Reddit bills itself as "the front page of the internet," operating as a source of both news and entertainment. Founded in 2005 and growing in users and subreddits ever since, Reddit is a flourishing online community. Data consists of more than 2.26 billion Reddit comments made from 2012 to 2016. I obtained Reddit comments from data dumps courtesy of pushshift.io (Baumgartner et al., 2020). I chose the period 2012 through 2016 as a five-year period beginning after a boom in Reddit usage from 2010-2011.

I limited the sample to only active subreddits—defined as subreddits attracting an average of at least 100 comments per month during their lifespan. I excluded subreddits that survived for only a single month, to exclude junk or spam subreddits. I also excluded non-English subreddits and subreddits dominated entirely by bots. At the comment level, I removed comments where the username or text had been deleted, removed, or did not exist. At the user level, I included only users who made at least 10 comments total in their lifetime on the site. For structural analyses, I also removed users who were likely to be bots. Additional details and explanations for these and other sampling criteria can be found in the S1 Appendix. The final sample includes a total of more than 2.26 billion comments, 13,757 unique subreddits, and more than 7.3 million unique users across 437,837 subreddit-month observations.

Measures

Dependent Variables.— The two dependent variables measure two different ways that subreddits can capture attention. Both variables represent rates of change in these modes of capturing

attention. The goal in measuring rates of change is to capture a subreddit's dynamic trajectory. Individual subreddits, and the overall ecology, are constantly shifting, and measuring rates of change helps to better capture these dynamics. The first dependent variable measures change in the size of subreddits in terms of the number of users—i.e. growth rate. This is a mode of success in which a subreddit captures a greater breadth of attention. Following many organizational studies (e.g., Geroski, 2005), I measure growth rate by dividing the size (i.e. number of users) of a subreddit in month *t* by the size in month *t-1* and then taking the natural log:

$$Growth_{it} = ln\left(\frac{NumUsers_{it}}{NumUsers_{it-1}}\right)$$

Where *i* denotes the subreddit of interest and *t* denotes the month. The second dependent variable measures the rate of change in the average number of comments per user in a subreddit. I call this measure "engagement." This is a mode of success in which a subreddit captures a greater depth of attention. I measure engagement in a similar way as growth rate:

$$Engagement_{it} = ln\left(\frac{NumCommentsPerUser_{it}}{NumCommentsPerUser_{it-1}}\right)$$

Measuring Structure.— The first key independent variable is structure. This variable is central to this analysis because it specifies the location of a subreddit in the overall resource space of users. To measure the structural location of subreddits, I use a weighted crowding measure. A standard structural crowding variable aggregates the resource overlaps between an entity and other entities in the ecology (e.g., Podolny et al., 1996; Wang et al., 2013)—here the user overlaps between a subreddit and other subreddits. In doing so, a crowding variable captures how connected a subreddit is with other subreddits on Reddit overall. Another way of stating this is that a structural crowding variable captures the amount of competition a subreddit faces for its users. A "crowded" location indicates high competition, whereas an uncrowded or "sparse" location indicates low competition.

I followed previous studies in organizational ecology (e.g., Podolny et al., 1996) in measuring crowding as the sum of niche overlaps. Niche overlap represents the proportion of shared users between two subreddits. For example, suppose there are two subreddits *i* and *j*. If subreddit *i* has 100 users, subreddit j has 50, and 25 users are members of both *i* and *j*, then *i*'s niche overlap value is .25, and *j*'s overlap value is .5. The structural crowding variable is constructed by adding up the overlaps between subreddit i and all other subreddits j in J that are not i.

To turn this standard crowding variable into a measure of position in the user space of Reddit, for each subreddit, I weight each user overlap by the size of the alter subreddit. This means that the crowding variable not only captures how crowded a subreddit's location is, but also how far this subreddit is from the largest subreddits on Reddit overall. A high structure score indicates that a subreddit overlaps with large, usually generic subreddits. In other words, subreddits with a high structure score are closely connected with the most prominent communities on Reddit. A low

structure score indicates that a subreddit does not have much overlap with the central locations on Reddit, and is thus relatively isolated. Overall, the structural crowding measure, which I refer to as structure for the sake of brevity, represents the sum of the weighted niche overlaps for the focal subreddit *i* with all other *j* subreddits at time *t*:

$$Structure_{it} = \sum_{i \neq j} ln(Size_{jt}) Overlap_{ijt}$$

Measuring Culture in Two Ways.— The second key independent variable is culture. I measure culture in two ways. These measures capture *what* kind of content a subreddit produces and thus what kind of taste it serves. Specifically, these measures capture the distinctiveness of content relative to the content of all other subreddits. This enables measurement of whether a subreddit serves a relatively generic or specialty taste. A high culture score indicates a specialty taste, whereas a low culture score indicates a generic taste. I refer to these measures as measures of culture for the sake of brevity.

I measure culture through the lens of language used in subreddits. Before calculating the measures, I conducted standard text-preprocessing steps on comment text data, including correcting spelling errors, removing hyperlinks, and lowercasing all words. I also ran an algorithm that detects common bigrams. For both measures, I included only words that appear at least 5 times in the entire Reddit comment data for each month, to limit the inclusion of junk words.

The first measure of culture is relatively simple, yet powerful. I follow Zhang et al. (2017) in measuring a subreddit's culture in terms of the distinctiveness of words appearing in that subreddit. Specifically, I measure the distinctiveness to subreddit *i* of each word *w* that appears in the subreddit, as the pointwise mutual information (PMI) between *w* and its subreddit context *i* relative to all subreddits *I*:

$$CulturePMI_{it} = \sum_{w} ln \frac{P_{it}(w)}{P_{It}(w)}$$

Where $P_{it}(w)$ is *w*'s frequency in the focal subreddit *i* at time *t* and $P_{It}(w)$ is *w*'s frequency across all subreddits I at time *t*. *w* is distinct to subreddit *i* if it occurs more frequently in *i* than in all subreddits *I*. These scores are calculated for each word occurrence *w* in i, and then logged and summed to produce a score for each subreddit *i* at each time *t*. Generic subreddits have a CulturePMI score close to 0, meaning words tend to appear just as frequently in that subreddit as in all other subreddits, while specialty subreddits have higher CulturePMI scores. See the S2 Appendix for example distributions of word PMI scores, as well as examples of words considered distinct and generic, within the context of particular subreddits.

I also measure culture in a more complex way that goes beyond relative word frequencies, to take the contextual meaning of words into account. To do so, I leverage a word embedding

model. Word embedding models generate vectors for words derived from word co-occurrences in the data. This generates an n-dimensional conceptual space, in which words that co-occur together frequently will be located close together, signifying semantic similarity. Word embedding models generate vectors for each word that capture multiple dimensions of meaning and thus multiple dimensions of similarity (Mikolov et al., 2013). In addition, word embeddings can capture similarities not only of words that co-occur with one another, but also that co-occur with shared context words.

I trained a word embedding model on each month of text data to produce 300-dimensional word vectors (results robust when using 100-dimensional vectors as well). The S3 Appendix provides additional detail on the specific model used and the training process. I then created a vector for every comment by averaging the word vectors within each comment. Following this, I created a vector for every subreddit by averaging the vectors of comments appearing in that subreddit in the given month.

Using these subreddit embeddings, I then calculated similarity scores. Within text embedding spaces, cosine-similarity is the preferred method of measuring similarity because it captures vector direction but not magnitude, and is thus independent of document length. To generate a culture measure for each subreddit, I measured the cosine similarity between that subreddit's vector and every other subreddit's vector in each month, and then calculated the average of these similarities. In doing so, I followed other studies that measured similarity across entities in embedding space (e.g., Burtch et al., 2022). This measure can be represented as follows:

$$CultureW2V_{it} = 1 - \frac{1}{J}\sum_{i \neq j} cos(v_{it}, v_{jt})$$

Where v_{it} is the vector representation of subreddit *i* at time *t*, $v_{jt}$is the vector representation of subreddit *j* at time *t*, and *cos* is the cosine similarity of these two vectors, defined as $cos(A, B) = \frac{AB}{||A||||B||}$. For subreddit *i* at time *t*, the CultureW2V score represents the average of all cosine similarities between v_{it} and v_{jt} for all other subreddits *j* in *J,* which I then subtract from 1 to obtain a measure of average cosine distance, such that a high score represents a specialty culture.

Controls.— In addition to the key independent variables, I also included several controls. First, I include a control for the age of the subreddit—measured as the number of months since the first comment in the subreddit. Second, during the time period of the study, Reddit administrators denoted some subreddits as "defaults"—meaning that new users automatically became members of these subreddits and thus were exposed to their content. Because "default" status is likely to have an impact on both user growth and engagement, I include a dummy variable denoting whether a given subreddit had default status during a given month. Third, I include a dummy variable for moderator activity. I captured whether each subreddit had comments removed by moderators—denoted in the raw data as comments where the text reads as "[removed]." I also

include the average length of comments in the subreddit (measured as the average number of characters per comment), in line with prior ecological studies of online communities (Wang et al., 2013).

Analysis

For the main analyses, I ran panel models for growth and engagement. I applied natural log transformations to all variables except age, moderator activity, and default status. Structure and culture variables were highly skewed, so I leveraged the lnskew0 command in Stata, which applies a log transformation on the variable plus or minus a constant chosen such that skewness is zero, and which allows log transformation of variables including zero values.

While many subreddits were active from the first until the last time they appeared, some subreddits have gaps, in which they became inactive and then active once again. Thus, the subreddit-months observed were a subset of the total possible subreddit-months. This represents a potential selection problem—endogeneity that could bias parameter estimates, as the reasons for being active or inactive in a given month are likely related to the outcomes of interest. In order to adjust for this selection issue, I ran a probit regression predicting the likelihood of a subreddit being active in a given month as a function of age and a random effect. I transformed the predicted probability of being active via inverse Mills ratio and included it as a control in the main growth and engagement models (Heckman, 1979; Tortoriello et al., 2012). Including this control did not significantly change parameter estimates in the main models. I specify the following model for growth and engagement:

$$
\begin{aligned}
y_{it} = {} & \beta_1 ln(NumUsers_{it-1}) \\
& + \beta_2 ln(CommentsPerUser_{it-1}) \\
& + \beta_3 ln(CommentLength_{it-1}) \\
& + \beta_4 ModActivity_{it-1} \\
& + \beta_5 Age_{it} \\
& + \beta_6 Default_{it-1} \\
& + \beta_7 ln(Culture_{it-1}) \times ln(Structure_{it-1}) \\
& + \beta_8 ProbabilityOfBeingActive_{it} \\
& + \gamma_i + \delta_t + \epsilon_{it}
\end{aligned}
$$

Where y_{it} is growth or engagement, γ_i is a subreddit fixed effect, δ_t is a month fixed effect, and ϵ_{it} is the idiosyncratic error term. Standard errors are clustered at the subreddit level.

Results

Distribution of Attention on Reddit

Before exploring the results from the growth and engagement models, I first show the distribution of attention to different subreddits. Figure 1 plots the log number of subreddits against the log number of comments in each subreddit to get a sense of the size distribution of

subreddits. This graph displays subreddits from 2016 that were included in the main sample for this article, from which many very small subreddits were removed. Even so, the plot reveals an extreme distribution of attention. The linear trend in log-log space suggests a scale-free distribution, indicating that a small number of subreddits account for a disproportionately large volume of comments. The S5 Appendix gives more detail on the different kinds of subreddits that attract varying amounts of attention.

[Insert Figure 1 about here]

Subreddit Growth

Table 1 reports results from models of subreddit growth. In line with the liability of newness, age has a positive effect on growth across all models, showing that older subreddits are better able to increase the size of their user base. The size of a subreddit (number of users) has a negative effect on growth across all models, suggesting that larger communities may struggle to continue growing. Subreddits that are assigned default status also grow more. Finally, subreddits where the average user comments more are able to grow more, however a longer average comment length has a negative effect on growth, suggesting a negative relationship between the depth of comments in a subreddit and growth. Mod activity has a non-significant effect across all models.

[Insert Table 1 about here]

[Insert Table 2 about here]

Models 2 and 3 consider the main effects of structure and the two culture measures. Across these models, structural crowding has a negative effect on growth, in line with much literature in organizational ecology (Carroll and Hannan, 1989). In these two models, a higher culture score (indicating a specialty culture) has a negative effect on growth. This suggests that a generic cultural profile is better for growth.

Models 4 and 5, the main models of interest, consider interaction effects between structure and culture. To more easily interpret these interaction effects, I examine the average marginal effects of the culture variable (average change in growth based on a one-unit increase in the culture measure) at different levels of structure. Table 2 displays these effects. The effect of the CulturePMI measure is very consistently negative across different levels of structure. For example, the effect at two standard deviations below the mean of structure is ($b = -0.0498$, $p < 0.001$) and the effect at two standard deviations above the mean of structure is ($b = -0.0327$, $p < 0.01$). For the CultureW2V measure, the effect is more contingent on structure. However, this effect is still negative except at the lowest values of structure, where culture has no significant effect. Culture here becomes more important at the highest structure levels, where there is the most competition for attention: ($b = -0.0400$, $p < 0.001$). These results suggest that subreddits with a more generic culture are better at attracting more users in structurally crowded areas. In less crowded areas, results are less consistent.

Subreddit Engagement

Table 3 reports results from models of subreddit engagement. Here again age has a positive, but very small effect, in line with the liability of newness. The size of communities has a positive effect on engagement across all models, showing that larger communities tend to attract deeper engagement. The average number of comments per user and average comment length have negative effects, suggesting that communities with already deep engagement struggle to attract deeper engagement. Being a default also has a negative effect on engagement. Mod activity has a non-significant effect across all models.

[Insert Table 3 about here]

[Insert Table 4 about here]

Models 2 and 3 consider the main effects of structure and the two culture measures. In these two models, a high structure score has a negative effect on engagement. A high culture score (indicating a specialty culture) has a positive effect on engagement. This suggests that a specialty culture is better for engagement. Models 4 and 5, the main models of interest, consider interaction effects between structure and culture. To more easily interpret these interaction effects, I again examine the average marginal effects of culture (average change in engagement based on a one-unit increase in the culture measure) at different levels of structure. Table 4 details these effects. For the CulturePMI measure, the effect of culture is slightly negative at lower levels of structure, but highly positive at higher levels of structure. For example, the effect at two standard deviations below the mean of structure is ($b = -0.0289$, $p < 0.01$) and the effect at two standard deviations above the mean of structure is ($b = 0.0936$, $p < 0.001$). For the CultureW2V measure, the effect of culture is non-significant at lower levels of structure and positive at higher levels of structure. For example, the effect at two standard deviations above the mean of structure is ($b = 0.0506$, $p < 0.001$). These results suggest that subreddits with a specialty culture are better at attracting deeper engagement from users in structurally crowded areas. In less crowded areas, results are less consistent, but generally suggest that culture may not matter as much in those locations. See the S6 and S7 appendices for robustness checks on these growth and engagement models with alternate measures of structure and engagement.

New Connections Online

The primary results of interest are summarized by the average marginal effects of the two culture measures on growth and engagement at different levels of structure, visualized in Figure 2. The results offer evidence that, in areas of high structural crowding—representing areas of connection and user overlap—both subreddits serving generic tastes and those serving specialty tastes can coexist by attracting attention in different ways: subreddits serving generic cultural tastes are better at attracting shallower engagement from more members, while subreddits serving specialty tastes are better at attracting deeper engagement from a smaller set of members.

These findings offer evidence of diverse cultural exploration at the individual level generating new connections between different kinds of communities at the ecological level online. In addition, these areas of high structural crowding reflect a community's overlap with large, generic communities and are thus the most popular and prominent locations in the overall ecology. The coexistence of generic and specialty communities here suggests that even new Reddit users, who readily gain exposure to a set of the most popular and generic communities, will encounter users who also belong to other communities representing a wide variety of tastes, facilitating cross-cutting discussion and opening pathways for exploration of diverse tastes.

[Insert Figure 2 about here]

At lower levels of structural crowding, communities are relatively isolated and independent, with less connection to other communities. The results suggest that the cultural profile of a community matters less here for how it attracts attention—individual communities can both grow and attract deep engagement, reflecting the fact that users here are dedicating their attention to one or very few communities.

Discussion

In this article, I investigated whether the internet reinforces patterns of cultural homogeneity in the offline world or instead facilitates cultural exploration of diverse, niche tastes. Using data from Reddit, I found evidence that both generic and specialty communities can capture the attention of the same individuals, reflecting diverse patterns of cultural exploration. While specialty communities may receive less attention than generic communities, they are able to occupy prominent locations, connected by shared users to large, generic communities. In contributing to the debate on the long tail of the internet, these findings reveal that while the overall distribution of attention to different kinds of communities may be similar online as offline, the connections between these communities—in terms of shared members—are vastly different in the online world. Underlying this finding is the idea that, on the internet, individuals can spread their attention across multiple and often diverse communities. The connections between generic and specialty communities resulting from these patterns of exploration have the potential to facilitate further cultural exploration through increased exposure of specialty communities catering to niche tastes.

This article offers evidence that the freedom of exploration at the individual level enabled by the internet loosens the coupling between structure and culture at the level of communities in the overall ecology. This finding adds to a body of literature in sociology suggesting the importance of studying structure and culture as related yet independent concepts, rather than assuming their coupling (e.g., Emirbayer and Goodwin, 2004; Mark, 2003). Specifically, I find that communities representing different cultural tastes can coexist even in the same structural location. This offers a contribution to studies of community and organizational ecology (e.g., Carroll, 1985; McPherson, 1983, 2004), and shows how ecological connections among communities can differ online. This difference is specifically undergirded by the different nature

of member resources online—individuals can serve as connection points between many communities, rather than existing in a single location in the resource space. Structural locations in the resource space do not map to discrete individuals pursuing those tastes, but instead to the fractions of peoples' attention allocated to those particular tastes.

Additionally, I find a consistent pattern in which cultural diversity flourishes in structurally crowded or competitive areas of the resource space, but not as much in sparse areas. This agrees with a long line of sociological theory dating back to Durkheim, who argued that social density operates as a form of competition producing an adaptive response in the form of differentiation, in which different units (e.g., individuals, communities, or organizations and institutions) in the system focus on different activities (Bellah, 1959: 452)—ideas which Fischer (1975) extended to explicitly connect social crowding with subcultural diversity. However, in the offline world, cultural differentiation in response to competition requires differentiation across the structural space of resources—such that different communities rely on different people (Carroll 1985; McPherson, 1983, 2004). On Reddit, I find evidence that communities serving different tastes can rely on the same individuals by capturing their attention in different ways. So while crowding in the offline world primarily reflects competition for resources, crowding in the online world can reflect sustainable connections across individuals and communities.

The contributions of this study are limited in several ways. First, this study is limited in measuring activity in online communities. Due to data availability, I was not able to capture more passive forms of engagement, such as views, clicks, and likes. While focusing on comments is powerful in that it centers the analysis on a relatively committed form of engagement—and thus shows where and how users allocate significant, rather than fleeting, attention—there may be different dynamics at play when looking at less active forms of content engagement. Second, this study was conducted on a single online platform. It may be that attention dynamics operate differently in other online settings, especially considering that ranking algorithms that drive attention to particular kinds of content may play a more prominent role on other sites. Different platforms have varied affordances that may interact with human attention in unique ways. Studying attention dynamics further on other platforms offers an exciting path for future research. However, the current findings from Reddit suggest that there are at least some online platforms that encourage cultural exploration and support the prominence of culturally diverse communities.

References

Adamic LA and Huberman BA (2000) Power-law distribution of the world wide web. Science 287(5461): 2115–2115.

Anderson C (2007) The Long Tail: How Endless Choice is Creating Unlimited Demand. New York: Random House.

Baldassare M (1992) Suburban communities. Annual Review of Sociology 18(1): 475–494.

Baumgartner J, Zannettou S, Keegan B, Squire M and Blackburn J (2020) The pushshift reddit dataset. In: Proceedings of the international AAAI conference on web and social media, volume 14. pp. 830–839.

Bellah RN (1959) Durkheim and history. American Sociological Review 24(4): 447–461.

Benkler Y and Nissenbaum H (2006) Commons-based peer production and virtue. Journal of Political Philosophy 14(4).

Bhattacharjee S, Gopal RD, Lertwachara K, Marsden JR and Telang R (2007) The effect of digital sharing technologies on music markets: A survival analysis of albums on ranking charts. Management Science 53(9): 1359–1374.

Breiger RL (1974) The Duality of Persons and Groups. Social Forces 53(2): 181–190.

Brynjolfsson E, Hu Y and Simester D (2011) Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. Management science 57(8): 1373–1386.

Burtch G, He Q, Hong Y and Lee D (2022) How do peer awards motivate creative content? Experimental evidence from Reddit. Management Science 68(5): 3488–3506.

Carroll GR (1985) Concentration and specialization: Dynamics of niche width in populations of organizations. American Journal of Sociology 90(6): 1262–1283.

Carroll GR, Dobrev SD and Swaminathan A (2002) Organizational processes of resource partitioning. Research in Organizational Behavior 24: 1–40.

Carroll GR and Hannan MT (1989) Density dependence in the evolution of populations of newspaper organizations. American Sociological Review 54(4): 524–541.

DiMaggio P, Hargittai E, Neuman WR and Robinson JP (2001) Social implications of the Internet. Annual Review of Sociology 27(1): 307–336.

Durkheim E (1984) The Division of Labour in Society. London: Macmillan Education UK. ISBN 978-0-333-33981-7 978-1-349-17729-5.

Elberse A and Oberholzer-Gee F (2006) Superstars and underdogs: An examination of the long tail phenomenon in video sales, volume 7. Harvard Business School, Boston: HBS Working Paper 07-015.

Emirbayer M and Goodwin J (1994) Network analysis, culture, and the problem of agency. American Journal of Sociology 99(6): 1411–1454.

Faraj S, Jarvenpaa SL and Majchrzak A (2011) Knowledge collaboration in online communities. Organization Science 22(5): 1224–1239.

Fischer CS (1975) Toward a Subcultural Theory of Urbanism. American Journal of Sociology 80(6): 1319–1341.

Fleder D and Hosanagar K (2009) Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. Management Science 55(5): 697–712.

Geroski PA (2005) Understanding the implications of empirical work on corporate growth rates. Managerial and Decision Economics 26(2): 129–138.

Goldhaber MH (1997) The attention economy and the net. First Monday 2(4 - 7).

Hannan MT and Freeman J (1989) Organizational ecology. Harvard university press.

Hawley AH (1971) Urban society: An ecological approach. (No Title) .

Heckman JJ (1979) Sample selection bias as a specification error. Econometrica 47(1): 153–161.

Johnson SL, Faraj S and Kudaravalli S (2014) Emergence of power laws in online communities. Mis Quarterly 38(3): 795–A13.

Lin Z, Salehi N, Yao B, Chen Y and Bernstein M (2017) Better when it was smaller? community content and behavior after massive growth. In: Proceedings of the International AAAI Conference on Web and Social Media, volume 11. pp. 132–141.

Macdonald D (1953) A theory of mass culture. Diogenes 1(3): 1–17.

Mark NP (2003) Culture and competition: Homophily and distancing explanations for cultural niches. American Sociological Review 68(3): 319–345.

McPherson M (1983) An Ecology of Affiliation. American Sociological Review 48(4): 519–532.

McPherson M (2004) A Blau space primer: prolegomenon to an ecology of affiliation. Industrial and Corporate Change 13(1): 263–280.

Mikolov T, Chen K, Corrado G and Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

Mohr JW and Guerra-Pearson F (2010) The duality of niche and form: The differentiation of institutional space in new york city, 1888–1917. In: Categories in markets: Origins and evolution. Emerald Group Publishing Limited, pp. 321–368.

Neuman WR (1991) The Future of the Mass Audience. Cambridge, UK: Cambridge University Press.

Newman ME (2003) The structure and function of complex networks. SIAM Review 45(2): 167–256.

Nicolaides B and Wiese A (2017) Suburbanization in the United States after 1945. In: Oxford research encyclopedia of American history.

Péli G and Nooteboom B (1999) Market partitioning and the geometry of the resource space. American Journal of Sociology 104(4): 1132–1153.

Peterson RA and Anand N (2004) The production of culture perspective. Annu. Rev. Sociol. 30: 311–334.

Peterson RA and Berger DG (1975) Cycles in symbol production: The case of popular music. American Sociological Review 40(2): 158–173.

Peterson RA and Kern RM (1996) Changing highbrow taste: From snob to omnivore. American sociological review : 900–907.

Podolny JM, Stuart TE and Hannan MT (1996) Networks, Knowledge, and Niches: Competition in the Worldwide Semiconductor Industry, 1984-1991. American Journal of Sociology 102(3): 659–689.

Salganik MJ, Dodds PS and Watts DJ (2006) Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. Science 311(5762): 854–856.

Shils E (1962) The theory of mass society: Prefatory remarks. Diogenes 10(39): 45–66.

Stigler GJ (1952) The Case Against Big Business". Fortune 44: 28–32.

Taeuscher K (2019) Uncertainty kills the long tail: Demand concentration in peer-to-peer marketplaces. Electronic Markets 29(4): 649–660.

Tan TF, Netessine S and Hitt L (2017) Is Tom Cruise threatened? An empirical study of the impact of product variety on demand concentration. Information Systems Research 28(3): 643–660.

TeBlunthuis N and Hill BM (2022) Identifying competition and mutualism between online groups. In: Proceedings of the international aaai conference on web and social media, volume 16. pp. 993–1004.

Tortoriello M, Reagans R and McEvily B (2012) Bridging the knowledge gap: The influence of strong ties, network cohesion, and network range on the transfer of knowledge between organizational units. Organization science 23(4): 1024–1039.

Waller I and Anderson A (2019) Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In: The World Wide Web Conference. pp. 1954–1964.

Wang X, Butler BS and Ren Y (2013) The Impact of Membership Overlap on Growth: An Ecological Competition View of Online Groups. Organization Science 24(2): 414–431.

Wilensky HL (1964) Mass society and mass culture: interdependence or independence? American Sociological Review 29(2): 173–197.

Zhang J, Hamilton W, Danescu-Niculescu-Mizil C, Jurafsky D and Leskovec J (2017) Community identity and user engagement in a multi-community landscape. In: Proceedings of the International AAAI Conference on Web and Social Media, volume 11. pp. 377–386.

Figures

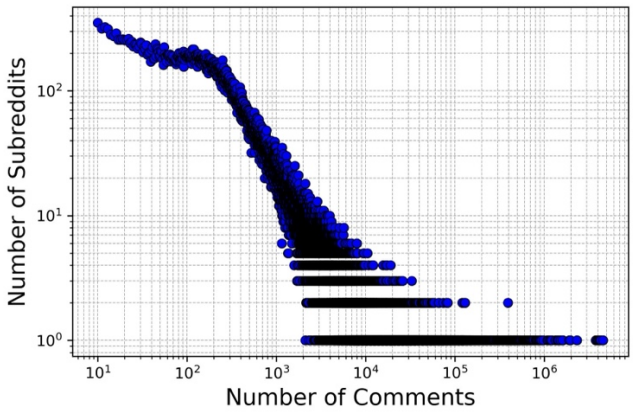

Figure 1: Log-log plot showing the distribution of the number of subreddits by their comment counts. Each point represents how many subreddits have exactly x comments.

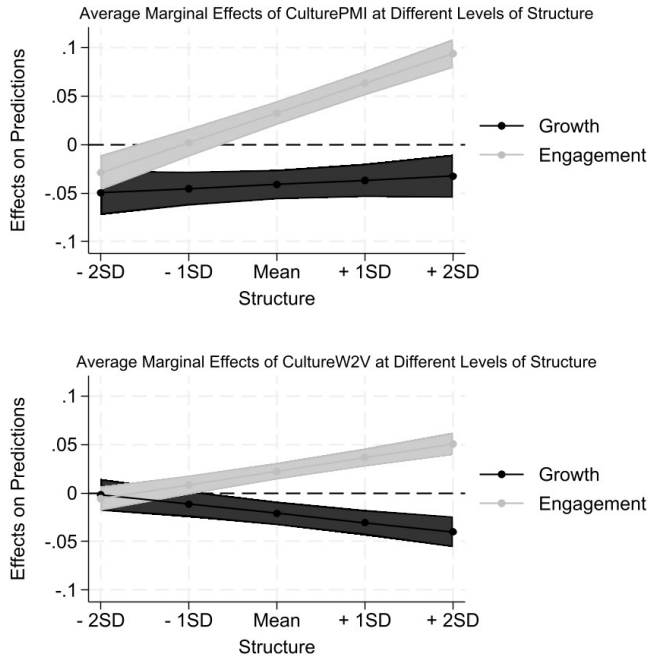

Figure 2: Visualizing Average Marginal Effects of Culture on

Tables

Table 1: Models Predicting Growth.

	(1)	(2)	(3)	(4)	(5)
Probability of Being Active	3.036***	3.028***	3.060***	3.020***	3.072***
	(0.347)	(0.347)	(0.347)	(0.347)	(0.347)
Num Users	-0.249***	-0.260***	-0.253***	-0.260***	-0.253***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Comments per User	0.097***	0.087***	0.093***	0.086***	0.094***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Comment Length	-0.041***	-0.048***	-0.043***	-0.048***	-0.043***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Default	0.196***	0.204***	0.201***	0.205***	0.203***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Age (Months)	0.005***	0.005***	0.005***	0.005***	0.005***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mod Activity	-0.006	-0.004	-0.004	-0.004	-0.004
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Structure		-0.119***	-0.118***	-0.123***	-0.220***
		(0.010)	(0.010)	(0.011)	(0.032)
CulturePMI		-0.040***		-0.118	
		(0.008)		(0.076)	
CultureW2V			-0.021***		0.149**
			(0.006)		(0.049)
Structure × CulturePMI				0.015	
				(0.015)	
Structure × CultureW2V					-0.034***
					(0.010)
Constant	1.085***	1.771***	1.633***	1.795***	2.141***
	(0.020)	(0.060)	(0.059)	(0.063)	(0.162)
Time FE	Yes	Yes	Yes	Yes	Yes
Subreddit FE	Yes	Yes	Yes	Yes	Yes
Observations	437,837	437,837	437,837	437,837	437,837
R-Squared	0.157	0.158	0.158	0.158	0.158

Notes: Models were estimated using cluster-robust standard errors at the level of the subreddit. * p<0.05, ** p<0.01, *** p<0.001

Table 2: Average Marginal Effects of Culture on Growth.

	CulturePMI	CultureW2V
-2SD	-0.0498***	-0.00204
	(0.012)	(0.008)
-1SD	-0.0455***	-0.0115
	(0.009)	(0.007)
Mean	-0.0412***	-0.0210***
	(0.008)	(0.006)
+1SD	-0.0370***	-0.0305***
	(0.009)	(0.007)
+2SD	-0.0327**	-0.0400***
	(0.011)	(0.008)

Notes: Table shows average marginal effects of culture measures across levels of structure. * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Table 3: Models Predicting Engagement.

	(1)	(2)	(3)	(4)	(5)
Probability of Being Active	5.519***	5.541***	5.508***	5.488***	5.489***
	(0.229)	(0.228)	(0.229)	(0.228)	(0.229)
Num Users	0.058***	0.066***	0.059***	0.066***	0.059***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Comments per User	-0.440***	-0.436***	-0.442***	-0.443***	-0.443***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)
Comment Length	-0.032***	-0.024***	-0.029***	-0.023***	-0.028***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Default	-0.094***	-0.098***	-0.095***	-0.096***	-0.098***
	(0.011)	(0.011)	(0.011)	(0.010)	(0.011)
Age (Months)	0.000***	0.000***	0.000*	0.000***	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mod Activity	0.000	-0.000	0.000	0.000	0.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Structure		-0.053***	-0.053***	-0.084***	0.098***
		(0.007)	(0.007)	(0.007)	(0.025)
CulturePMI		0.039***		-0.516***	
		(0.006)		(0.052)	
CultureW2V			0.023***		-0.231***
			(0.004)		(0.038)
Structure x CulturePMI				0.109***	
				(0.010)	
Structure x CultureW2V					0.051***
					(0.008)
Constant	0.244***	0.440***	0.579***	0.609***	-0.178
	(0.014)	(0.042)	(0.039)	(0.042)	(0.129)
Time FE	Yes	Yes	Yes	Yes	Yes
Subreddit FE	Yes	Yes	Yes	Yes	Yes
Observations	437,837	437,837	437,837	437,837	437,837
R-Squared	0.215	0.216	0.216	0.217	0.216

Notes: Models were estimated using cluster-robust standard errors, at the level of the subreddit. * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Table 4: Average Marginal Effects of Culture on Engagement.

	CulturePMI	CultureW2V
-2SD	-0.0289**	-0.00590
	(0.009)	(0.006)
-1SD	0.00172	0.00823
	(0.007)	(0.005)
Mean	0.0323***	0.0224***
	(0.006)	(0.004)
+1SD	0.0630***	0.0365***
	(0.006)	(0.005)
+2SD	0.0936***	0.0506***
	(0.008)	(0.006)

Notes: Table shows average marginal effects of culture measures across levels of structure.

* $p<0.05$, ** $p<0.01$, *** $p<0.001$