Attention Allocation in Online Communities

James C. Mellody

January 21, 2024

Abstract

Online communities are increasingly important sites of social and economic life. These communities depend on the attention of member resources in order to survive. This article asks the question: how is attention allocated in online communities? Early work highlighted a key promise of the internet: to counteract homogenizing cultural forces in society and provide a more egalitarian context for a greater diversity of content to find an audience. A large body of literature has found instead that the allocation of attention online resembles resource allocation in many offline settings: characterized by power-law distributions in which a few offerings attract the bulk of attention. Moving beyond just the distribution of attention, I undertake an ecological analysis of online communities, measuring the location of communities both in a structural resource space of users as well as in a cultural topic space based on language. In line with past work, I find that the most generic communities attract the bulk of attention. However, my study identifies a new pattern: by examining community location in the ecology, I find that niche communities are not relegated to the periphery of the ecology, but rather can succeed in the very center, right next to the most dominant, generic communities. This suggests a significant difference in how the allocation of attention occurs online, and a means through which online platforms provide space for culturally diverse communities to flourish.

1 Introduction

Life—social, civic, economic, and organizational—increasingly occurs online. Much of social life flows through online channels, where individuals connect with friends and strangers alike around shared interests to build virtual communities (Ren et al. 2012). Beyond daily connection, online communities offer spaces for peer production, resulting in paradigm-shifting developments like Wikipedia and Linux (Stewart 2005, Fitzgerald 2006), as well as platforms for crowdsourced innovation, problem-solving (Lifshitz-Assaf 2018), and funding (Peterson and Wu 2021), facilitating the development of new solutions to key societal and economic problems and funding the latest entrepreneurial ventures on platforms like Kickstarter and GoFundMe. Online communities also offer spaces for collective civic action, most notably catalyzing social movements as seen in widespread protests facilitated by digital media such as the Arab Spring, Black Lives Matter, and Occupy Wall Street (Castells 2015, Caren et al. 2020). The reach of online communities extends even to traditional firms, offering means for employees to engage with customers (Fisher 2019), share knowledge (Neeley and Leonardi 2018, Mickeler et al. 2023), and build community (Porter and Donthu 2008), increasingly important with the spread of remote and hybrid work.

Like voluntary organizations, social movements, and traditional firms before them, online communities depend on building an audience and attracting individuals willing to contribute. Online communities run on attention—depending ultimately on individuals as member resources who both consume and produce content. The network of relationships between individuals and the communities they contribute to comprises an online ecology of affiliation—a digital parallel to previously described ecologies of affiliation between voluntary organizations and their members (Breiger 1974, McPherson 1983, McPherson and Rotolo 1996, McPherson and Smith-Lovin 2002). Given that attention is a limited resource vital for the survival of online communities (Goldhaber 1997), a key question is: how is attention allocated in online communities?

Early work painted the internet as the great attention equalizer: a countervailing force against cultural homogeneity (Neuman 1991, DiMaggio et al. 2001, Anderson 2007). In many offline markets, resource distributions follow a power law, with a few generic offerings attracting the bulk of resources (Anderson 2007, Brynjolfsson et al. 2011), and forming a market center, while niche offerings attract small audiences, and survive only in the periphery of the market (Carroll 1985, Carroll et al. 2002). The dominance of generic offerings is an issue of unmet demand—with high search costs and readily available generic offerings, individuals often settle for offerings that are good enough, but which do not fully meet their needs.

A key promise of the internet was to reduce the amount of people forced to settle for readily-accessible, generic options, and instead to enable people to more easily optimize against their preferences (Brynjolfsson et al. 2011). Specifically, the internet lowers search costs and barriers to entry (Brynjolfsson et al. 2011, Faraj et al. 2011). Online, trying something new requires only clicking a few links or typing a few lines of text—inexpensive actions relative to the hefty temporal and economic costs of trial and error in the offline world. In addition, the internet offers nearly unlimited and relatively costless space for a variety of offerings to proliferate. For online communities, this signaled a more even distribution of attention—a move away from power-law distributions in which a few generic options dominate and toward a flatter distribution in which a long-tail of niche offerings could attract significant attention. Some research focused on online marketplaces found initial empirical evidence in support of this shift (Anderson 2007, Bhattacharjee et al. 2007, Brynjolfsson et al. 2011, Dewan and Ramaprasad 2012, Zentner et al. 2013, Zhang 2018).

Yet a large body of research since the beginning of the 21st century has found that the allocation of attention online resembles familiar patterns of resource allocation in offline markets. In contrast to the long-tail of niche offerings, some work found instead the reproduction of "superstar" or "rich-get-richer" effects online (Elberse and Oberholzer-Gee 2007, Fleder and Hosanagar 2009, Hosanagar et al. 2014, Tan et al. 2017, Park et al. 2020). Work more explicitly examining the allocation of attention in online communities has found striking similarities to offline markets: competitive crowding (Wang et al. 2013, Zhu et al. 2014), scale-based advantages and competitive differentiation of community form (Lin et al. 2017, Hwang and Foote 2021), and power-law distributions of attention (Newman 2003, Barabási 2009, Johnson et al. 2014). A key explanation for the emergence of familiar patterns is that attention is still a limited resource online (Goldfaber 2017), and that individuals rely on readily available information about the choices of their peers when making their own decisions (Barabási and Albert 1999, Salganik et al. 2006). Given these familiar patterns of attention allocation, the promise of greater success for a diversity of offerings online may be an empty one, with the reality of the online world instead reflecting the same patterns of homogeneity and dominance of generic offerings that describe many offline contexts.

However, most existing work studying attention allocation online takes a distributional perspective, meaning this work examines how much of the total share of attention different kinds of communities attract (e.g., Brynjolfsson et al. 2011, Wang et al. 2013). A key issue is that this approach overlooks how the "resource space" is fundamentally different in online communities. A concept from organizational ecology, "resource space" denotes the social-structural density map of available resources (i.e. users for online communities), with the center representing the area of greatest resource abundance, and the periphery representing areas of relative resource scarcity (Péli and Nooteboom 1999). Power law distributions emerge in contexts with high search costs, and reflect resources that are discrete and resource allocation that is zero-sum. Different kinds of offerings must rely on entirely separate resource pools. In online communities, where search costs are low, individual resources can serve multiple communities in equilibrium (Waller and Anderson 2019, Teblunthuis et al. 2022), such that resources are no longer discrete and resource allocation no longer zero-sum. This means that different offerings (here communities) can potentially rely on the same individuals as resources in equilibrium—by capturing their attention in different ways.

In order to investigate the difference this makes in the allocation of attention in online communities, I move beyond a purely distributional examination of attention allocation in three ways. First, drawing on organizational ecology, I locate online communities within the social-structural space of user resources, thereby enabling an examination of *where* different kinds of communities can succeed in the overall ecology. Second, I independently measure the location of online communities in a cultural space based on language—capturing *what* kind of taste a community targets (i.e., generic or niche). This goes beyond existing ecological studies of organizations and online communities that assume a coupling between cultural taste and location in the structural resource space (i.e. generic communities in the center and niche communities in the periphery), and allows for decoupling between structure and culture as a result of users being able to participate in multiple communities. Finally, I examine *how* communities succeed, allowing for the possibility that attention can be captured in different ways, rather than only discretely and in a zero-sum fashion. I do this by measuring two success outcomes for communities: growth (increase in users) and engagement (increase in the average number of comments made by each user).

In a study of the online community Reddit, a massive online platform consisting of thousands of topic-specific sub-communities, I find similarities to existing work in terms of the relationship between *what* kind of taste a community serves and *how* it can succeed. In line with power-law distributional patterns, as well as patterns in ecological studies of markets (Carroll 1985), I find that it is the most generic communities that can grow the largest, while more niche communities rely on attracting deeper engagement from fewer, more dedicated individuals. However, I identify a key difference in attention allocation online in terms of *where* different kinds of communities can succeed: namely that online, both generic and niche communities can find success anywhere in the ecology. This means that, rather than being relegated to the periphery of the ecology where only a few users exist (Carroll et al. 2002), niche communities can find success in the most prominent locations in the ecology, right next to the largest, most generic communities. This suggests a difference in how the allocation of attention occurs online, and highlights a key way in which the internet can bolster the prominence of a diversity of cultural offerings and activities.

More specifically, the location of niche communities next to generic communities reflects a high-level of user overlap, meaning that many of the very same individuals engage in both kinds of communities. This means greater connectivity and cross-pollination among different kinds of communities, and signals a breakdown of traditional barriers between resource bases for different kinds of content. While this diverse set of niche communities may not gain the bulk of attention, it can exist online in closer proximity to the most prominent communities, offering greater exposure and potential for cross-cutting discussion and engagement. These findings, and this study's emphasis on examining the success of different kinds of communities based on their position in a structural network of users, also offer a contribution to the conversation on whether the internet results in a "global village" or instead a "cyber-Balkans" (Van Alystne and Brynjolfsson 2005), meaning greater unification or fragmentation online. Specifically, these findings suggest that niche communities need not be relegated to the periphery where they are reliant on isolated users fragmented from the main user population. Rather, niche communities can coexist alongside generic communities, connected by shared users.

This article is organized as follows: first, I review existing theory on the promise of the internet; second, I examine several ways in which the reality of the internet instead reflects familiar patterns of resource allocation offline; third, I detail a different approach to studying attention allocation that informs the empirical analyses; I then introduce the setting and data, and detail the method of analysis; after this I conduct a deep exploratory analysis of Reddit; I then present results from models examining how different kinds of Reddit communities can find success in different locations in the ecology of the platform; I conclude with a discussion of the contributions of this article to our understanding of the long tail phenomenon, fragmentation versus unification online, and organizational ecology.

2 Theory

2.1 The Promise of the Internet

The 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). In this respect, the internet initially appeared to be a ready-made solution to a key societal and economic issue: the homogenization of culture.

Discussions of cultural homogenization took off with mass-culture theory (e.g., MacDonald 1953, Wilensky 1964), portending the homogenizing force of mass-media and increasingly national markets and institutions, and continued into the later 20th century with studies of market structure. These analyses found that increasingly concentrated markets, consisting of corporations making broad-based appeals to common-denominator tastes in an attempt to draw in large audiences, led to a "massification" of tastes (Shils 1962). While specialty offerings could survive, they found success only during short-lived periods of increased market competition (Peterson and Berger 1975, Peterson and Anand 2004), or were otherwise permanently relegated to the periphery of the market, surviving off of a small base of highly engaged consumers (Carroll 1985). Similarly, access to communities inclusive of diverse interests and social identities proved difficult outside of the largest, most densely populated urban areas (Fischer 1975)—a problem of increasing concern with widespread suburbanization in the 20th century (Baldassare 1992). Overall, these issues concern resource availability—many niche offerings struggle to find the resources necessary to survive, while generic, mass-produced offerings dominate based on their broad appeal.

The internet promised to level the playing field, offering the space and freedom for individuals to explore a diversity of tastes—not only by consuming products through online marketplaces, but also by finding spaces for discussion and co-production with like-minded others in online communities (Neuman 1991). Geographic constraints would no longer limit individuals—including those historically disenfranchised—from finding communities of like-minded others (Wellman 2001). And individuals who previously settled for generic cultural offerings in the absence of readily accessible alternatives could now easily locate options better matching their specific tastes. Scholars hypothesized that these shifts would generate a more balanced distribution of attention across a diversity of products and communities. Much of this work considers online marketplaces, and argues that the digital transition results in more evenly distributed resources among so-called "superstar" offerings and a "long-tail" of niche offerings (Anderson 2007, Brynjolfsson et al. 2011, Zentner et al. 2013, Zhang 2018).

Several key characteristics of the internet fueled these predictions. Overall, the internet increases freedom of choice and lowers formal barriers present in more traditional settings. First, it lowers costs of search and evaluation, making it easy for individuals to try out many different options and optimize against their preferences (Brynjolfsson et al. 2011). In online communities, individuals can evaluate whether a particular space matches their preferences with only a few clicks of a mouse and a few minutes of reading. Second, online communities have fluid boundaries and low membership barriers, which enable individuals to easily become active participants and contributors to communities (Faraj et al. 2011). Third, the internet lowers costs of production, storage, and distribution—offering nearly limitless space for a variety of options to coexist, including content made in a grass-roots or peer-produced fashion (Benkler and Nissenbaum 2006, Fitzgerald 2006).

2.2 Familiar Patterns in the Allocation of Attention Online

Despite these hopes and predictions, a large body of research paints a picture of the internet that strays from the promise of cultural diversity. 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).

Research has identified three key familiar patterns. First, competition for resources hinders growth. Several studies locate individual online communities in a larger ecology of communities, finding that communities whose users and topics overlap more directly with other communities struggle to grow as large as communities with more moderate levels of overlap (Wang et al. 2013, Zhu et al. 2014). The main idea is that resources are limited, and so communities that rely on resources that many other communities rely on face greater competition for the time and attention of those resources.

Second, there is evidence of community differentiation and associated scale-based sources of advantage, through which different kinds of online communities find divergent paths to success. This reflects a differentiated base of users who participate in online communities for a variety of reasons (Wasko and Faraj 2005) and who seek out different kinds of communities to fulfill different needs and wants (Teblunthuis et al. 2022, Waller and Anderson 2019). Within an ecology of communities, this means that in addition to large, generalist communities, there is commonly a proliferation of smaller, specialized communities (Hwang and Foote 2001). While larger communities may benefit from economies of scale and scope in producing a greater volume of content (i.e., requiring fewer contributions per user to do so), smaller communities can offer a different kind of shared experience (Lin et al. 2017). These patterns are reminiscent of how consumers in traditional markets select on organizational form identity, such that generalist and specialist organizations can coexist in partitioned markets by serving different tastes (Carroll 1985, Carroll and Swaminathan 2000).

Third, the overall distribution of attention online resembles resource distributions in the offline world. Namely, there is evidence that the allocation of attention online follows a power-law distribution (Newman 2003, Johnson et al. 2014). In fact, power laws describe many online dynamics, from user contributions (Johnson et al. 2015), to the popularity of particular websites (Adamic and Huberman 2000), to the degree distribution of user networks and sales in online marketplaces (Elberse and Oberholzer-Gee 2007, Tan et al. 2017, Taeuscher 2019). These power-law distributions point to "superstar" effects and winner-take-all dynamics—in which a few offerings capture the bulk of attention. This is reminiscent of the dominance of large, generalist firms in many traditional product markets (Carroll 1985).

Explanations for these patterns revolve around the central idea that attention remains a limited resource online, and that individuals make choices in how to allocate their attention (Goldhaber 1997). The presence of competitive crowding based on user overlap suggests limits to attention, and the coexistence of large, generalist communities and small, specialist communities suggests competitive differentiation. Similarly, power-law distributions of attention suggest that individuals pay attention to the choices of others in deciding how to allocate attention. At the same time that an online context makes it easier for individuals to explore options, it also offers more direct information on peer preferences, which can bolster social influence effects leading to herding behavior (Adamic and Huberman 2000, Aral and Walker 2014). Through mechanisms such as preferential attachment, individuals seek out other individuals and communities that are already well-connected (Barabási and Albert 1999), or, through mechanisms of social influence and status signaling, engage with content already approved by their peers (Salganik et al. 2006).

Overall, these familiar patterns of attention allocation suggest that the internet may not be the bastion of egalitarianism and diversity many initially hoped, with the reality of the online world instead reflecting similar patterns of homogeneity and dominance of generic offerings that describe many offline contexts. Was the promise of the internet an empty one?

2.3 A Different View of Resource Allocation Online

Despite many familiar patterns emerging in the allocation of attention online, there is evidence of a key difference at the individual level. To appreciate this difference, it is first important to illustrate how allocation works at the individual level offline. In offline settings, where it is often difficult to search across the space of offerings, individuals tend to restrict their activity to a cluster of related offerings (Zuckerman and Kim 2003, Sine et al. 2005, McKendrick and Hannan 2014), optimizing as much as they can to find "the products that best match their preferences" (Péli and Nooteboom 1999: 1133). This means that an individual can only serve as a resource for a single or small set of related organizations or offerings. Individuals are discrete resources which organizations compete for in a zero-sum fashion (Carroll et al. 2002). This has implications for organizations at the market level: serving generic or niche tastes requires relying on entirely different resource bases, and thus occupying different positions in the overall market (Carroll and Hannan 1989, Swaminathan 1995, Carroll and Swaminathan 2000, Dobrev et al. 2002, Negro et al. 2014). Generic offerings exist in the center of the market consisting of many people holding popular tastes, and specialty offerings exist in the periphery consisting of a separate and smaller pool of people with more niche tastes. There is little overlap between these two resource bases, relegating cultural diversity to the periphery of the market.

Online, these dynamics are different. Specifically, many individuals spread their attention quite broadly in online communities (Waller and Anderson 2019, Teblunthuis et al. 2022). The low search costs, low barriers to entry, and fluid boundaries of online communities mean it is possible for individuals to concurrently belong to multiple, potentially diverse communities. While multiple membership often represents an unsustainable drain on an individual's time and attention in offline membership ecologies (Zald and Ash 1966, McPherson 1983), it is much more sustainable in online communities, and thus a widespread phenomenon. Empirical evidence suggests that many online community users participate across multiple communities—comprising "packages" or "suites" of communities (Teblunthuis et al. 2022; Teblunthuis and Hill 2022). These differences also have implications for communities at the ecological level. No longer are resources discrete entities captured in a zero-sum fashion by organizations competing in a market; instead, online communities can capture the attention of individuals in different, fractional ways. This means that different online communities can potentially rely on the same resources in equilibrium.

Examining solely the distribution of attention to different communities—the focus of much existing work—may obscure how the allocation of attention to online communities reflects these differences. In order to move beyond a purely distributional examination of attention allocation, I independently measure *where* communities exist in the overall resource space of users, *what* kind of cultural taste they target, as well as *how* they find success. First, I locate online communities within the social structural space of users. To do so, I measure structural location based on user overlaps—users that overlap between different communities (e.g., Podolny et al. 1996, Wang et al. 2013). This situates each community in the overall resource space—specifically measuring whether a community is in a relatively crowded area and thus faces higher competition for user attention, or whether the community is in a relatively sparsely populated area and does not face as much competition for user attention. Second, leveraging multiple measures of culture, I capture the cultural taste that a community targets. Whereas in many offline settings, the cultural taste an organization targets is assumed to be coupled with its position in the structural research space (Carroll and Swaminathan 2000), here I allow for the possibility that culture can be decoupled from position in the structural resource space—reflecting the idea that different communities can potentially rely on the same resources in equilibrium. Finally, I measure multiple modes of success, following the idea that attracting a greater volume of attention is not the only way to succeed. Specifically, I examine success both in terms of growth in the number of users, and also increasing engagement, measured as the average number of comments per user.

3 Data and Empirical Strategy

3.1 Data

Data for this study comes 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. There are generic subreddits around large topics such as "r/games," which covers any and everything game related, as well as more specialized subreddits such as "r/babyelephantgifs," which, as advertised, focuses on sourcing GIFs of baby elephants.

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 2.25 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. 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. Additional details and explanations for these and other sampling criteria can be found in Appendix A. The final sample includes a total of more than 2.25 billion comments, 14,212 unique subreddits, and more than 7.3 million unique users across 440,235 subreddit-month observations.

3.2 Measures

3.2.1 Dependent Variables.

Following many studies of organizational growth rate (e.g., Geroski 2005, Chen et al. 2012), I measure the first dependent variable, 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. I measure the second dependent variable, engagement, as the average number of comments per user made in the focal subreddit. I measure change in engagement in the same way as growth rate:

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

3.2.2 Measuring Structure

The first key independent variable is structure. This variable is key to this analysis because it specifies the location of a subreddit in the overall resource space of users. Specifically, we need a measure that captures whether a subreddit is located in a relatively competitive or noncompetitive space—signifying, respectively, centers of the resource space versus peripheral areas. I do not restrict the resource space to a single center, but rather, in order to examine the relative competitiveness of a subreddit's location, allow for the possibility of many "centers" of competition in the space.

Structural crowding offers a suitable measure. 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. However, I adapt this measure by weighting niche overlaps by the size of the alter subreddit. This means that the structure measure here captures not only competitiveness, but also distance from the largest, most prominent communities. Continuing the example, then, I weight each niche overlap by the logged size of the alter subreddit. For subreddit i, this means multiplying .25 by ln(50), which equals 0.978. For subreddit j, this means multiplying 0.5 by ln(100), which equals 2.303. Overall, structural crowding, 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 in time t:

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

3.2.3 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 targets. Specifically, these measures capture the distinctiveness of content relative to the content of all other subreddits. This enables measurement of whether a subreddit is targeting a relatively generic or niche taste. A high culture score indicates a niche taste, whereas a low culture score indicates a generic taste. I refer to these measures as measures of culture, rather than cultural distinctiveness, for the sake of brevity.

In measuring culture, I follow recent work that has examined culture through the lens of language (e.g., Srivastava and Goldberg 2017, Srivastava et al. 2018, Kozlowski et al. 2019). I similarly 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—meaning that, for example, occurrences of "New York" in the data are treated as such rather than as "new" and "york" separately. For both measures, I included only words that appear at least 5 times in the entire Reddit comment data for each month, to eliminate garbage 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 s of each word w that appears in the subreddit, as the pointwise mutual information (PMI) between w and its subreddit context s relative to all subreddits S:

$$CulturePMI_{st} = \sum_{w} ln \frac{P_{st}(w)}{P_{St}(w)}$$

Where $P_{st}(w)$ is w's frequency in the focal subreddit s at time t and $P_{St}(w)$ is w's frequency across all subreddits S at time t. w is distinct to subreddit s if it occurs more frequently in s than in all subreddits S. These scores are calculated for each word occurrence w in s, and then logged and summed to produce a score for each subreddit s 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 distinct subreddits have higher CulturePMI scores. See Appendix B 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 meanings 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 (Aceves and Evans 2023). Word embedding models generate vectors for each word that capture multiple dimensions of meaning and thus multiple dimensions of similarity. This allows for more complex associations such as the classic example of word algebra: vector("king") − vector("man") + vector("woman") results in a vector closest to vector("queen") (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. A key advantage of word embedding models is that they allow for comparison between words, and aggregated up, entire corpuses, that takes into account greater complexity of semantic relationships, grounded in the meaning of the words as used in the corpus itself.

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). Appendix C provides additional detail on the specific model used and the trainings 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. To generate a culture measure for each subreddit, I measured the cosine similarity between that subreddit's vector and every other subreddit's vector, 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. 2021; Carlson 2022; Guzman and Li 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 distinctive culture.

3.2.4 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 included the average length of comments in the subreddit (measured as the average number of words per comment), in line with prior ecological studies of online communities (Wang et al. 2012).

3.3 Analysis

For the main analyses, I ran panel fixed effects 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 (Casciaro and Piskorski 2005; Zhu and Chen 2014). Results are robust when models are run with standard log transformations.

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.

4 Preliminary Descriptive Exploration of Reddit

4.1 Key Correlations

Table 1 shows descriptive results for untransformed variables, and Table 2 shows the correlation matrix for transformed variables. I explore some key correlations in detail below.

——————————— Insert Table 1 about here ———————————

——————————— Insert Table 2 about here ———————————

Size is a very important variable in the consideration of organizational ecologies. Existing theory argues that size lines up both with market position and the level of cultural distinctiveness of the content an organization produces, with large generalists producing mass-market products and occupying the densest, most competitive areas, and small specialist producing niche products and occupying the sparse peripheries (Carroll 1985). I examine here whether these relationships hold at a descriptive level in the Reddit data.

First, I consider the relationship between culture and size. For the CulturePMI measure, there is a strong negative relationship between culture and size—showing that the most culturally distinct subreddits are often small in size, while the most culturally generic subreddits are often large. This makes sense: the larger a subreddit is, the harder it is to sustain a unique language focused on an area that does not relate to other subreddits. In other words, as subreddits grow, the influx of new users from other areas means that the content tends to become more generic overall. This suggests some similarities between offline and online patterns in terms of size and culture. Considering the CultureW2V measure, there is a still negative, yet not nearly as strong, relationship with size. Given that this measure is not strictly dependent on word frequencies, and instead dependent on word meaning in context, it makes sense that there is a weaker relationship between culture and size.

——————————— Insert Figure 3 about here ———————————

Finally, as seen in Figure 3, there is no strong relationship between size and the structural position of subreddits. There is variation all along the size spectrum in terms of the structural position of subreddits, suggesting that subreddits of different sizes occupy a variety of structural positions—a key difference from many offline contexts.

4.2 Power-Law Distribution of Attention on Reddit

Before exploring the structure and culture of subreddits, I first show the distribution of attention to different subreddits. Figure 1 below displays a plot of the number of comments made on Reddit, by subreddit. The plot reveals an extreme distribution of attention, with a small fraction of subreddits receiving the bulk of attention. This distribution is even more extreme than resource distributions in offline settings. In Figure 2, I binned the subreddits by comment size, to give a better sense of the distribution. Overall, these two plots show that a familiar, power-law distribution of resources occurs on Reddit.

——————————— Insert Figure 1 about here ———————————

——————————— Insert Figure 2 about here ———————————

4.3 Visualizing Structure and Culture

Figure 4 below displays two graphs visualizing the structure and culture of Reddit in December 2016—the last period in the data included in this article. For structure, I created 300-dimensional vectors of each subreddit based on user membership, doing so by calculating a truncated singular value decomposition on a subreddit-by-subreddit matrix of user overlaps.[1] For culture, I leverage the subreddit embeddings underlying the CultureW2V score.

——————————— Insert Figure 4 about here ———————————

[1]See Robustness Checks for additional information on how I created these embeddings.

I created the visualizations below using the t-SNE procedure to plot in two-dimensions the 300-dimensional vectors of structure and culture (Van der Maaten and Hinton 2008).[2] In the graph, each circle denotes a subreddit, and the size of the circle reflects the number of users in the subreddit during that month.

The different colors in the structural graph represent different clusters of subreddits based on overlapping users, identified by running a K-means clustering algorithm on the subreddit structure vectors. The structural graph shows that there is not just one singular market center, but rather multiple clusters of subreddits with many overlapping users. There is one cluster consisting of the largest subreddits. Many of these subreddits include the most general, high level subreddits, such as r/AskReddit, r/funny, r/pics, r/videos, and r/gifs. However, there is also an abundance of smaller subreddits that exist in positions very close to these largest subreddits. There are also additional clusters of subreddits—both dense and sparse and consisting of large and small subreddits—in other parts of the space. In other words, the picture here does not resemble a traditional partitioned market with a single market center and a periphery consisting of small organizations. This graph also helps to show the meaning of the structure variable. Subreddits in dense areas, indicating many overlapping users, will have high structure scores. The highest structure scores will be seen in those subreddits that have overlapping users with the largest subreddits, indicating high competition for users.

I also include a graph that maps subreddits based on their cultural content. I created this graph from the subreddit culture vectors underlying the CultureW2V measures. I again mapped these into a two-dimensional representation using the t-SNE procedure. And once again, here the different colors in the graph represent different clusters of subreddits based on cultural similarity, identified by running a K-means clustering algorithm on the subreddit culture vectors. The size of the circles here again reflects the size of the subreddits. The first point to note is that culture offers a related, yet distinct picture of the subreddit space. This underscores the idea that structure and culture are related, yet distinct elements of social connection (Emirbayer and Goodwin 1994), and thus two ways of mapping the location of communities in space. Another important point is that there is a great diversity of cultural content—represented by numerous clusters of content, rather than a single center and then only diversity at the periphery.

Finally, in Figure 5, I include the same structural graph as in the previous figure. Here, however, the colors represent CulturePMI scores. Darker shades indicate higher CultureW2V scores, meaning more distinctive cultures, whereas lighter shades indicate more generic cultures. What is important to highlight here is that we can see distinct subreddits and generic subreddits overlapping and occupying locations right next to one another. This offers descriptive evidence of a key difference in resource allocation online—specifically that different kinds of communities can rely on the same users and occupy the same locations in the structural resource space. This illustrates the importance of measuring structure and culture independently.

——————— Insert Figure 5 about here ———————

[2] One point to note is that t-SNE visualizations preserve local structure more accurately than global structure. I thus introduce this graph for exploratory purposes, rather than to draw final conclusions.

4.4 Variation in Subreddit Structure and Culture

In this section I provide examples of subreddits in different structural positions and with different cultural profiles. Below in Figure 6 is a graph of the four quadrants (separated by the means of the structure and culture variables) with the examples of specific subreddits located based on their average monthly values of structure and culture (CulturePMI). This offers further descriptive evidence of a decoupling between structure and culture, and the question I will investigate in the main empirical section is whether different cultural profiles can find success in these different structural locations.

——————————— Insert Figure 6 about here ———————————

First I consider the lower left quadrant—containing subreddits with both low structure and low culture scores. There are two kinds of subreddits that fit this profile. The first is very large generalist subreddits that have effectively outgrown competition. Basically these subreddits are ecosystems on their own, monopolizing the attention of users. These subreddits are where many new users first go when they join Reddit. One example is r/IAmA. r/IAmA is a very large subreddit, where people can start a thread by declaring "I am" followed by their profession or a unique experience they have had, and then people in the comments will ask questions about their life and work—a process known as an "ask me anything" session. Many celebrities have posted in this subreddit, often as a promotion for an ongoing project (e.g., movie, album, tv show). Most famously, former President Barack Obama sat for an AMA session in 2012, actually crashing the site briefly. This community has an average of 79,815 users making a total of 194,749 comments per month. The other profile in this quadrant consists of smaller but still self-contained communities, like those for a particular city or community, such as r/Charlotte—for Charlotte, NC. To give a sense of the scale difference, r/Charlotte has an average of 851 users making 33,373 comments per month. Content here is relatively generic as well—content that could apply to living in any city such as a recent post looking for "Summer sport activities for [my] kid."

Next, the upper left quadrant contains subreddits with low structure but high culture scores. These are subreddits that are isolated both structurally and culturally. These are very niche subreddits populated by users with distinct interests, who do not participate across many other subreddits. These subreddits tend to be very small. Examples include subreddits focused on obscure video games or TV shows, or obscure offshoots of more mainstream content. A specific example of a subreddit in this quadrant is r/ClubNintendoSwap, which has an average of 50 users making a total of 175 comments per month. While Nintendo is a hugely popular video game company, Club Nintendo was a niche customer loyalty program, where customers could obtain rewards for providing feedback on Nintendo products—rewards that they could in turn exchange for limited edition items such as trading cards. The purpose of r/ClubNintendoSwap was to create a space for users to "trade club Nintendo rewards"—effectively a marketplace for those seeking to swap rewards. Another example is r/HomeKit, for discussion of the Apple Home app, a software platform for interfacing with smart home devices. This subreddit had an average of 60 users making a total of 161 comments per month.

The final two quadrants occupy structurally crowded areas—subreddits here have relatively high structure scores and so share many of their users with other subreddits, especially the largest subreddits. The bottom right quadrant contains subreddits with low culture scores. Examples include quite popular and large subreddits that deal with generic content but are relatively less self-contained, such that they often slot into a larger suite of subreddits engaged in by their users. These include subreddits like r/LosAngelesRams, for fans of the NFL team, which has an average of 866 users making a total of 4,607 comments per month. Another example is the r/Liberal community, a generic subreddit for discussion of liberal leaning politics, which has an average of 578 users making a total of 1,867 comments per month.

Finally, the upper right quadrant contains subreddits with high structure scores and high culture scores. These are subreddits which produce relatively distinct cultural content but are still closely connected structurally to many other subreddits, including the largest subreddits. These subreddits are usually smaller in size but still occupy the densest areas of the resource space. Examples include subreddits that combine popular and niche content, or niche content that fits within a broader suite of participation such that these subreddits do not function as self-contained spaces where only isolated users gather. Specific examples include subreddits focused on niche humor, like r/GoodFakeTexts, which focuses on discussion of funny screenshots of fake text exchanges, and which has an average of 153 users making a total of 231 comments on average per month. Another example is r/MapsWithoutNZ—a subreddit dedicated to chronicling the surprisingly common and amusing phenomenon of maps that unintentionally omit New Zealand (including, the top post of all time in the subreddit, a map from New Zealand's own official government web page). This subreddit has an average of 215 users per month making a total of 302 comments per month. If these subreddits can find a way to succeed, this suggests that even in densely populated, competitive areas, subreddits targeting relatively "niche" or "esoteric" tastes can thrive, meaning that generic tastes do not completely dominate.

5 Main Analyses

5.1 Growth

Table 3 reports results from models of subreddit growth. Model 1 considers the effects of key controls, without the main variables of interest, structure and culture. 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. When success is considered in terms of pure quantity of users, then the more developed content of older communities may be beneficial. 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 3 about here ———————————

——————————— Insert Table 4 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 distinctive 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.[3] For the PMI Culture measure, this interaction is not significant, and there remains a negative main effect of culture. For the W2V Culture measure, this interaction is significant.

To more easily interpret these effects, I examine the average marginal effect of culture at different levels of structure (meaning different locations in the resource space of users). Additionally, I examine predicted outcomes at key levels of both structure and culture. Table 4 contains average marginal effects of culture at different levels of structure. Culture has a consistent negative effect across different levels of structure. The effect of the PMI culture measure is very consistent across different levels of structure. For example, the effect at two standard deviations below the mean of structure is (β = -0.0479, $p < 0.001$) and the effect at two standard deviations above the mean of structure is (β = -0.0436, $p < 0.001$). For the W2V culture measure, the effect is more contingent on structure. However, this effect is still negative except at the lowest values of structure. Culture here becomes more important at the highest structure levels, where there is the most competition for attention: (β = -0.0310, $p < 0.001$). Figures 7 and 8 provide visualizations of these effects, showing predicted values at different key values of structure and culture. The most important takeaway is the consistent negative effect of culture across the different levels of structure, and across two measures of culture.

——————————— Insert Figure 7 about here ———————————

——————————— Insert Figure 8 about here ———————————

5.2 Engagement

Table 5 reports results from models of subreddit engagement. Model 1 considers the effects of key controls, without the main variables of interest, structure and culture. 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, suggesting that larger communities tend to attract deeper engagement. The average number of

[3]Running models with squared structure terms, and a three-way interaction between structure squared and culture, yields consistent results, but I omit this squared structure term in models reported here for the sake of interpretability. I do the same for engagement models below.

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 5 about here ———————————

——————————— Insert Table 6 about here ———————————

Models 2 and 3 consider the main effects of structure and the two culture measures. Across these models, a high structure score has a positive effect on engagement. In these two models, a high culture score (indicating a distinctive culture) has a positive effect on engagement. This suggests that a distinctive cultural profile is better for engagement. Models 4 and 5, the main models of interest, consider interaction effects between structure and culture. For the PMI Culture measure, this interaction is not significant, and there remains a positive main effect of culture. For the W2V Culture measure, this interaction is significant.

To more easily interpret these effects, I again examine the average marginal effect of culture at different levels of structure, as well as predicted outcomes at key levels of both structure and culture. Table 6 contains average marginal effects of culture at different levels of structure. Culture has a consistent positive effect across different levels of structure. The effect of the PMI culture measure is very consistent across different levels of structure. For example, the effect at two standard deviations below the mean of structure is (β = 0.0556, p < 0.001) and the effect at two standard deviations above the mean of structure is a bit stronger at (β = 0.0637, p < 0.001). For the W2V culture measure, the effect is more contingent on structure. However, this effect is still consistently positive across all values of structure. For example, the effect at two standard deviations below the mean of structure is (β = 0.0243, p < 0.001) and the effect at two standard deviations above the mean of structure is a bit stronger at (β = 0.0460, p < 0.001). Figures 9 and 10 provide visualizations of these effects, showing predicted values at different key values of structure and culture. The most important takeaway is the consistent positive effect of culture across the different levels of structure, and across two measures of culture. What is most unusual here is that subreddits can survive with a distinct cultural profile, even in the most competitive structural locations of the resource space.

——————————— Insert Figure 9 about here ———————————

——————————— Insert Figure 10 about here ———————————

5.3 Tradeoff and Cultural Diversity

Taken together, the results suggest that subreddits face a tradeoff: target a generic cultural taste to attract more users and grow, or target a niche cultural taste to attract deeper engagement from users in the form of more comments per user. These findings map on to power law distributions of attention—it is the few, generic offerings that attract the bulk of attention from users, while a plethora of small offerings attract deeper engagement from a smaller user base.

However, the results of the analyses above offer something new in terms of *where* different kinds of offerings can survive. By measuring structure and culture independently, I examined where in the resource space different kinds of cultural offerings succeeded. I found that the tradeoff described in the previous paragraph operates across all locations in the resource space, such that both generic and niche offerings can find success in any location in the space of users. This is a key difference from how resources are allocated in offline markets, in which different kinds of offerings are located in different locations in the market. On Reddit, even very niche offerings can find success in the most prominent, crowded locations in the resource space, right next to the largest, most generic communities. Thus, while it may not be that attention is distributed more evenly away from generic offerings and toward more niche offerings in terms of pure volume, it is the case that niche offerings can succeed in a more prominent location in the overall resource space, meaning that there is a breakdown in the separation of resources consuming different kinds of content.

5.4 Robustness Checks

To examine the robustness of results, I conducted several additional tests. First, I ran models for growth and engagement with a different measure of structure. I do not use this measure in the main analyses because it can capture higher order user overlaps, rather than only direct user overlaps between two subreddits, which is what I want to measure. However, the differences in this measure offer a nice check of the robustness of the results to alternative means of measuring structure.

This is the same structure measure underlying the structure t-SNE plots of Figures 4 and 5. To construct this measure, I followed a process detailed in Martin's community2vec paper (2017). The overall idea is, instead of a crowding variable, to create a vector representation of each subreddit based on their user memberships. First, I created a subreddit-by-subreddit matrix, in which cells denoted the number of overlapping users between subreddits, for all possible subreddit pairs. I then transformed this matrix by calculating the positive pointwise mutual information of each cell. Each subreddit thus had a sparse vector representation, of thousands of dimensions, each cell representing its overlap with another subreddit, and with many cells being zero due to no overlap. In order to transform this sparse vector into a dense vector of a smaller dimension, I then computed the truncated singular value decomposition of this matrix (Hamilton et al. 2016). The truncated singular value decomposition is a dimensionality reduction technique which takes the largest k singular values of a matrix that explain the most variance. I chose k to be 300, producing a dense 300-dimensional vector for each subreddit, denoting the subreddit's structural position based on its overlapping users with other subreddits. I transformed this into a cosine similarity score, following the same procedure as for the CultureW2v measure. In doing so, however, I weighted each cosine similarity score by the logged size of the alter subreddit, in order to here again make this into a measure of distance from the largest subreddits, and thus a measure of competitive position in the structural resource space.

Full results are shown in Appendix D. Results are largely consistent with the main models. However, there are some key differences. One difference is that, in the growth models, in the areas of structural sparsity, rather than there being no difference between generic and distinct cultures, it is actually better to be distinct in order to grow. For engagement models, it is still better across all structural locations to be culturally distinct, although the effect of culture (for the CulturePMI measure) is more contingent on structural location. Despite these differences, the results are qualitatively consistent overall, with generic cultures favoring growth and distinct cultures favoring engagement. The consistencies across these two different measures of structure, as well as across two different measures of culture, offer evidence of the robustness of the results.

Next, I ran engagement models with a different measure of engagement. Specifically, I measured engagement as change in the average length of comments made in a subreddit, measured in words. Whereas the measure used in the main analyses counts the number of comments per user overall, here this measure gets at engagement in a slightly different way, specifically whether users are engaging more deeply by writing longer, and thus more in-depth, comments. Results are again largely consistent, yet once again different in a similar way to the last robustness check using a different structure measure. In models with the CultureW2V measure, there is no significant effect of culture on engagement at lower levels of structure. See Appendix E for full results.

6 Discussion

The advent and increasing prominence of the internet yielded a question: would the internet offer space for a diversity of niche offerings to flourish, or instead reinforce the dominance of generic, culturally homogenous offerings? This article makes progress on resolving this question in two key ways. First, this article confirms past findings that the internet has, in some cases, reinforced power-law distributions of resources, and offers some additional insight into why this may be the case. On Reddit, I find that a small proportion of communities attract the bulk of attention, and that there is a similar relationship as in many offline markets between *what* kind of content a community produces and *how* it can succeed. Specifically, it is a generic set of communities that can grow large, reinforcing the dominance of homogeneity in the overall resource distribution, while niche communities can attract deeper engagement from a smaller set of users. This is in line with findings in organizational ecology of offline markets showing the dominance of generic firms, while niche specialty firms must rely on a small set of enthusiast consumers (Carroll 1985, Carroll and Swaminathan 2000).

I argue that this finding can be explained by two similarities between online and offline contexts. First, offerings—whether organizations offline or communities online—compete for limited resources. Online, attention remains a limited resource (Goldhaber 1997). Second, individuals must choose how to allocate their limited attention. In doing so, individuals use information available to them in order to facilitate their choices. The choices that others make offer valuable information about quality of various offerings, providing a pathway for social influence and the emergence of power-law distributions of attention allocation (e.g., Adamic and Huberman 2000, Aral and Walker 2014). These factors help to account for the similarities we see in how resources are allocated offline and online, specifically in the form of competitive dynamics revolving around scarce resources: competitive crowding (Wang et al. 2013, Zhu et al. 2014), scale-based advantages and competitive differentiation of community form (Lin et al. 2017; Hwang and Foote 2021), and power-law distributions of attention (Newman 2003, Barabási 2009, Johnson et al. 2014).

The second, and primary, way in which this article makes progress on resolving the question posed above, is to provide evidence of a key difference in how attention is allocated online, thereby suggesting a means through which the internet provides a space for cultural diversity to flourish. I do this by moving beyond a distributional analysis of attention. I find that there is a key difference from offline contexts in terms of *where* different kinds of communities can succeed in the overall ecology. On Reddit, both generic and niche communities can survive anywhere in the overall ecology. They do so by relying on the same resources in different ways. In terms of the whether the internet allows diverse content to flourish, this suggests that while niche communities may not receive more attention online, they are able to occupy spaces of increasing prominence, right next to the largest, most generic communities. This is in stark contrast to resource allocation in offline contexts, where specialist offerings and firms are relegated to peripheral market locations (Carroll 1985, Carroll and Swaminathan 2000).

Online, niche communities can gain greater exposure through cross-pollination with the most popular communities. This suggests that the average Reddit user, who initially joins large, generic communities, will easily and naturally be able to stumble upon a diverse set of niche communities. This also suggests that conversations can flow across generic and niche communities, because of the many overlapping users. Diverse perspectives and spaces can thus become more prominent and influential in the overall ecology. In addition to the long-tail literature, this finding offers a contribution to the related debate on whether the internet will lead to "global village" or instead a "cyber-Balkans" (Van Alystne and Brynjolfsson 2005), meaning greater unification or fragmentation online. These findings suggest that niche communities need not rely on niche users, but that instead the very same users can connect generic and niche content. The idea that people are pursuing fragmented tastes, such that each community is populated by separate user bases, is not reflected in the findings here. Instead, the findings suggest a more interconnected network, in which different kinds of communities exist in close proximity to one another, connected by shared users.

The finding of increasing prominence for niche communities is rooted in differences between offline and online contexts. Online spaces offer lower costs and boundaries (Brynjolfsson et al. 2011). Previous studies, however, have found mixed results in terms of whether these changes actually increase allocation of attention to diverse offerings (Brynjolfsson et al. 2010). To address this, I focus on an understudied difference at the individual level: that people can engage with multiple communities concurrently (Waller and Anderson 2019, Teblunthuis et al. 2022). At the level of the ecology, this means that the very nature of the resource space is different—no longer are resources discrete, and competition for resources zero-sum. Instead, online communities can capture the attention of the same individuals, but in different ways. This means that different kinds of communities can potentially rely on the same resources—I find empirical support for this on Reddit, where both generic and niche subreddits are able to succeed alongside one another in the resource space, by capturing the attention of users in different ways.

Related to these points, these findings offer contributions to our understanding of organizational ecology and resource partitioning. While past work has shown that crowding can impact growth in online spaces, just as it can in offline markets (Wang et al. 2013), the present article shows that online contexts can support different kinds of communities even in the same location in the resource space. Underlying this is a difference in the nature of the resource space online versus offline. In offline contexts, the resource space resembles a density map of discrete users, with each occupying a single position. In choosing to target a particular cultural taste, organizations must locate in the section of the resource spaces where resources pursuing that taste exist. This is because resources allocate their attention in a relatively discrete manner due to the costs of evaluating different offerings. In online communities, the resource space is different. Because individuals can allocate their attention across multiple communities, we can conceptualize the resource space instead as a social network rather than a map of separate resource pools. This goes back to the idea that the network of communities and users is an online ecology of affiliation (McPherson 1983), representing a duality between individuals and communities (Breiger 1974). Community nodes exist in localized pockets of crowding or sparsity defined by their overlapping members with other communities. There are potentially numerous such pockets, rather than a bifurcated market consisting of a center and periphery, as seen in Figure 4.

This understanding of the resource space as a social network also helps to make sense of some additional findings from the present study. Specifically, there is a consistent pattern across several different specifications that shows cultural differentiation matters more in crowded areas compared to less crowded areas. Seen through a social network lens, areas of higher crowding are not just areas of greater competition for users, but also areas of greater availability of attention. This means that these more crowded areas can better support a greater diversity of content, because the users there are spreading their attention more broadly. In offline contexts, crowding similarly means competition, but because resources are captured in a discrete fashion, crowding does not mean greater resource availability to support cultural diversity. Cultural differentiation is an adaptive response to competition (Bellah 1959, Fischer 1975), but cultural diversity can only flourish where there is resource availability. In offline contexts, cultural diversity is relegated to the periphery, where there are pockets of unmet demand. Cultural differentiation necessitates structural differentiation. Online, where resources can be spread across multiple offerings, cultural differentiation can occur without structural differentiation.

Building from this, and at a more meta-theoretical level, the present article contributes to ecological theory, and echoes a growing body of literature in sociology, by suggesting the importance of studying structure and culture as related yet independent concepts, rather than assuming their coupling (Emirbayer and Goodwin 1994, Ruef 2000, Mark 2003). An assumption of coupling makes sense in offline contexts where high costs couple individual resources with a particular taste or set of related tastes (Peterson and Berger 1975, Carroll 1985). However, in online communities, where individuals can engage with potentially numerous cultural tastes simultaneously, communities can differentiate culturally even in the same structural resource location, capturing individuals' attention in different ways, and meaning that structure and culture can become decoupled.

Overall, the findings in this article suggest that the internet may reinforce certain familiar patterns of resource allocation, but also can open up new avenues for diverse content to find an audience and enter into the larger conversation. While this study was limited to a single online platform, studying these dynamics on other platforms with different affordances offers an exciting path for future research.

7 Tables

Table 1: Descriptive Statistics

	Mean	SD	Min	Max
Growth	1.177	2.524	0.00182	624.8
Engagement	1.044	1.145	0.00284	409.4
Structure	172.2	234.0	0	15085.5
CulturePMI	1.022	0.613	0.00271	14.42
CultureW2V	0.0434	0.0397	0.00304	0.718
Num Users	998.5	8169.6	1	583739
Comments per User	5.239	22.89	1	1901.6
Comment Length	36.19	21.98	1	892.1
Age (Months)	39.20	25.38	0	131
Default	0.00487	0.0696	0	1
Mod Activity	0.241	0.428	0	1

Table 2: Correlation Table

	1	2	3	4	5	6	7	8	9	10	11
(1) Growth	1										
(2) Engagement	0.348	1									
(3) Structure	-0.0289	-0.0142	1								
(4) Culture_PMI	0.114	0.0743	0.0491	1							
(5) Culture_W2V	0.00603	0.00569	0.102	0.531	1						
(6) Num Users	-0.144	-0.0310	-0.0118	-0.770	-0.198	1					
(7) Comments per User	-0.0707	-0.209	-0.0518	-0.257	0.133	0.132	1				
(8) Comment Length	-0.00724	-0.0136	-0.0628	-0.360	-0.222	0.00526	0.128	1			
(9) Age	-0.0225	0.0220	0.0521	-0.318	-0.0529	0.385	-0.103	0.169	1		
(10) Default	-0.00168	0.000380	-0.00334	-0.172	-0.0267	0.231	-0.0142	0.00261	0.0874	1	
(11) Mod Activity	-0.0561	-0.0235	0.0803	-0.132	0.215	0.227	0.123	-0.00611	0.276	0.0252	1

Table 3: Growth Models

	(1)	(2)	(3)	(4)	(5)
Probability of Being Active	3.196***	3.142***	3.185***	3.145***	3.164***
	(0.346)	(0.345)	(0.345)	(0.345)	(0.345)
Num Users	-0.248***	-0.259***	-0.250***	-0.259***	-0.250***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Comments per User	0.091***	0.088***	0.095***	0.088***	0.095***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Comment Length	-0.043***	-0.049***	-0.043***	-0.049***	-0.042***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Default	0.192***	0.196***	0.192***	0.196***	0.192***
	(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	-0.005	-0.003	-0.004	-0.003	-0.004
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Structure		-0.024***	-0.025***	-0.025***	-0.053***
		(0.002)	(0.002)	(0.002)	(0.011)
CulturePMI		-0.045***		-0.054*	
		(0.007)		(0.024)	
CultureW2V			-0.020***		0.027
			(0.006)		(0.018)
Structure × CulturePMI				0.002	
				(0.005)	
Structure × CultureW2V					-0.009**
					(0.003)
Constant	1.089***	1.283***	1.141***	1.286***	1.284***
	(0.020)	(0.026)	(0.028)	(0.026)	(0.061)
Time FE	Yes	Yes	Yes	Yes	Yes
Subreddit FE	Yes	Yes	Yes	Yes	Yes
Observations	440,235	440,235	440,235	440,235	440,235

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: Growth Margins

	CulturePMI	CultureW2V
-2SD	-0.0479***	-0.00697
	(0.010)	(0.008)
-1SD	-0.0468***	-0.0130*
	(0.008)	(0.007)
Mean	-0.0457***	-0.0190**
	(0.007)	(0.006)
+1SD	-0.0447***	-0.0250***
	(0.008)	(0.006)
+2SD	-0.0436***	-0.0310***
	(0.009)	(0.007)

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

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

Table 5: Engagement Models

	(1)	(2)	(3)	(4)	(5)
Probability of Being Active	5.265***	5.314***	5.252***	5.320***	5.271***
	(0.229)	(0.228)	(0.228)	(0.227)	(0.228)
Num Users	0.056***	0.071***	0.059***	0.071***	0.059***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Comments per User	-0.434***	-0.426***	-0.436***	-0.426***	-0.436***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Comment Length	-0.023***	-0.013**	-0.020***	-0.013**	-0.020***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Default	-0.086***	-0.094***	-0.090***	-0.094***	-0.090***
	(0.011)	(0.011)	(0.011)	(0.011)	(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	0.001	-0.000	0.000	-0.000	0.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Structure		0.008***	0.008***	0.007***	0.034***
		(0.002)	(0.002)	(0.002)	(0.010)
CulturePMI		0.060***		0.044*	
		(0.006)		(0.020)	
CultureW2V			0.036***		-0.006
			(0.005)		(0.016)
Structure × CulturePMI				0.003	
				(0.004)	
Structure × CultureW2V					0.008**
					(0.003)
Constant	0.221***	0.071**	0.285***	0.076***	0.156**
	(0.014)	(0.022)	(0.019)	(0.022)	(0.054)
Time FE	Yes	Yes	Yes	Yes	Yes
Subreddit FE	Yes	Yes	Yes	Yes	Yes
Observations	440,235	440,235	440,235	440,235	440,235

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 6: Engagement Margins

	CulturePMI	CultureW2V
-2SD	0.0556***	0.0243***
	(0.008)	(0.006)
-1SD	0.0576***	0.0297***
	(0.007)	(0.005)
Mean	0.0596***	0.0352***
	(0.006)	(0.005)
+1SD	0.0617***	0.0406***
	(0.007)	(0.005)
+2SD	0.0637***	0.0460***
	(0.008)	(0.006)

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

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

8 Figures

Figure 1: Power Law Distribution of Attention

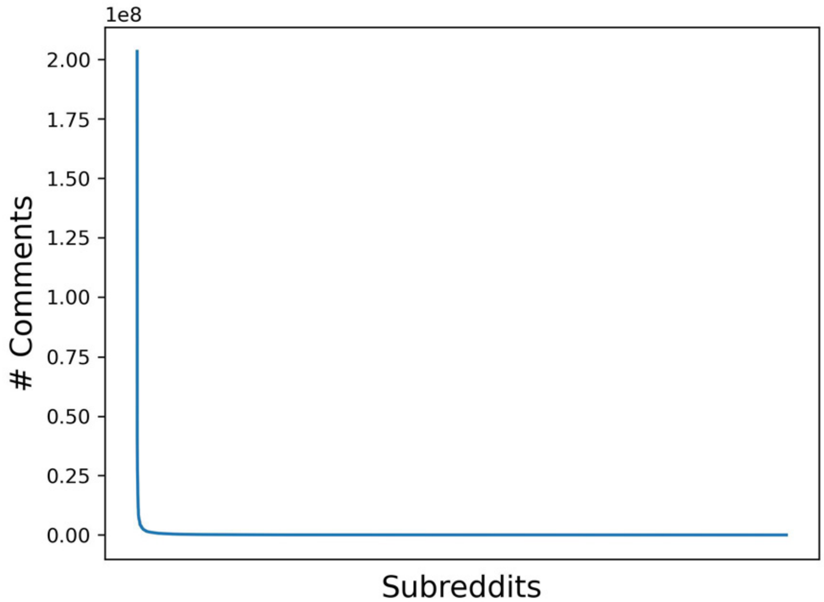


Figure 2: Subreddits Binned by Size

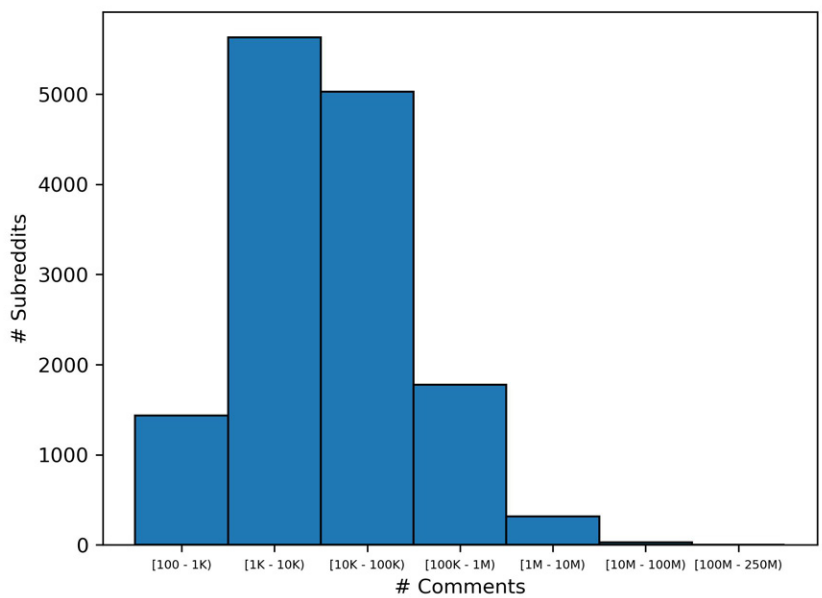


Figure 3: Relationship between Key Measures and Size

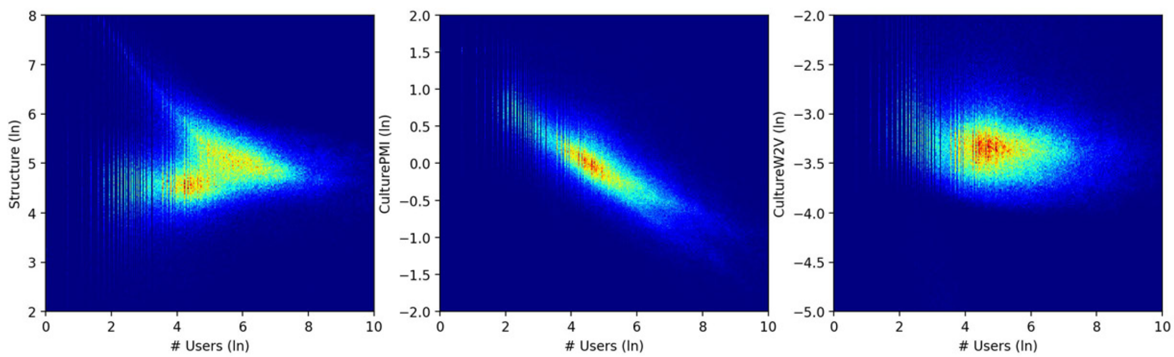


Figure 4: Visualizing Structure and Culture

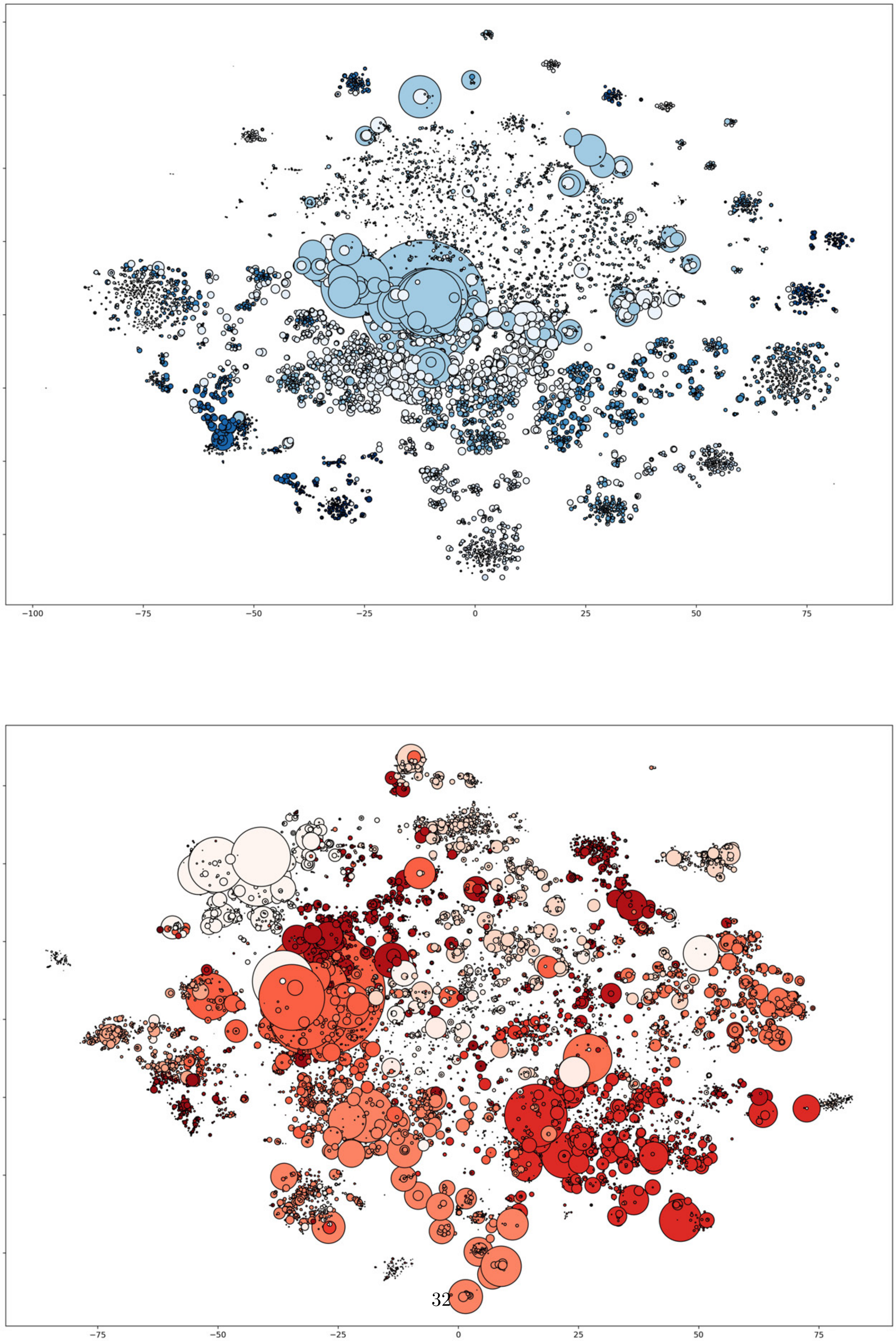


Figure 5: Culture within Structure

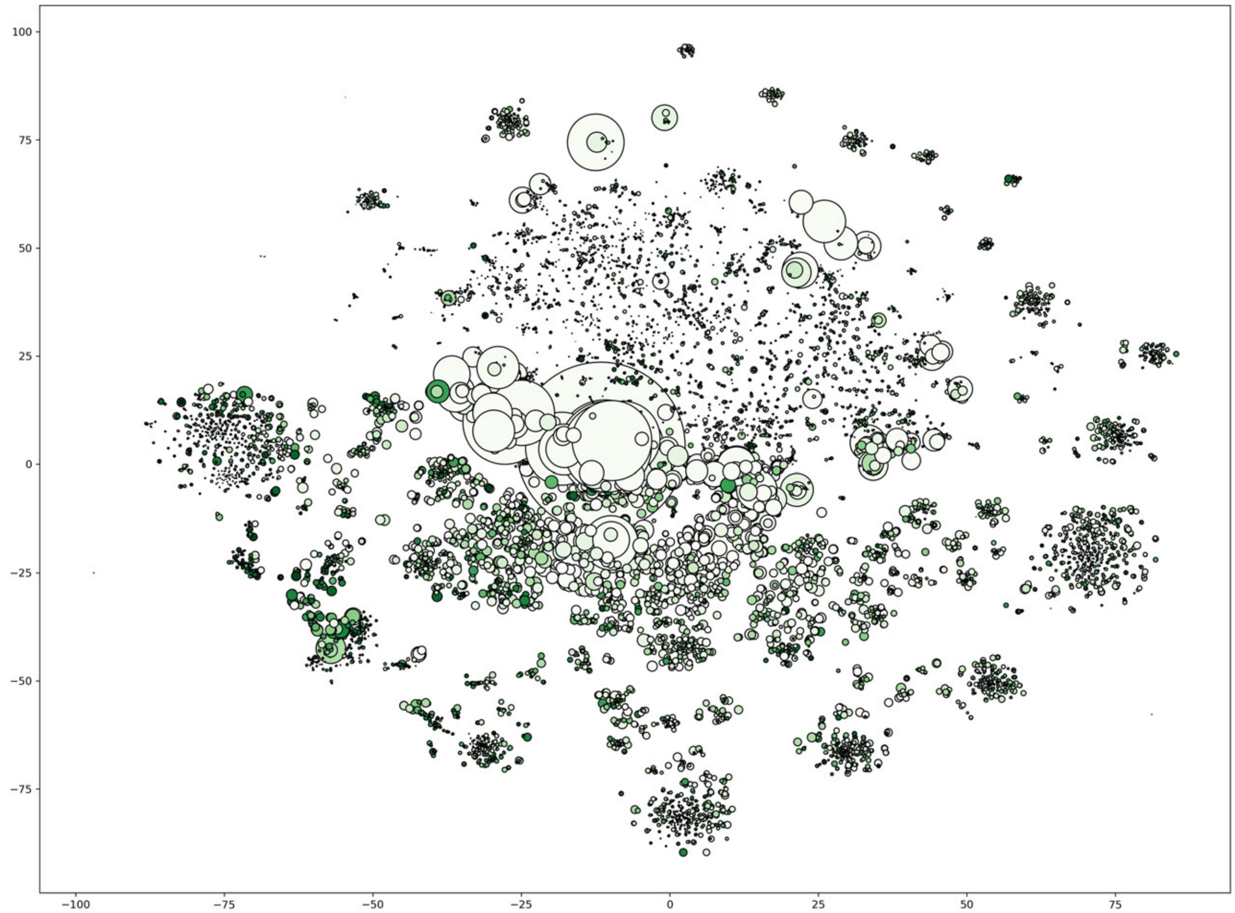


Figure 6: 4 Quadrants

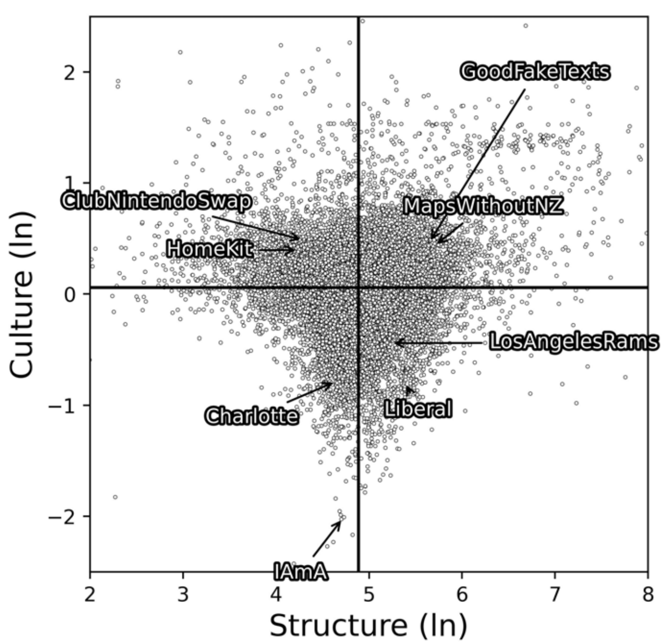


Figure 7: Predicted Growth Outcomes - CulturePMI Measure

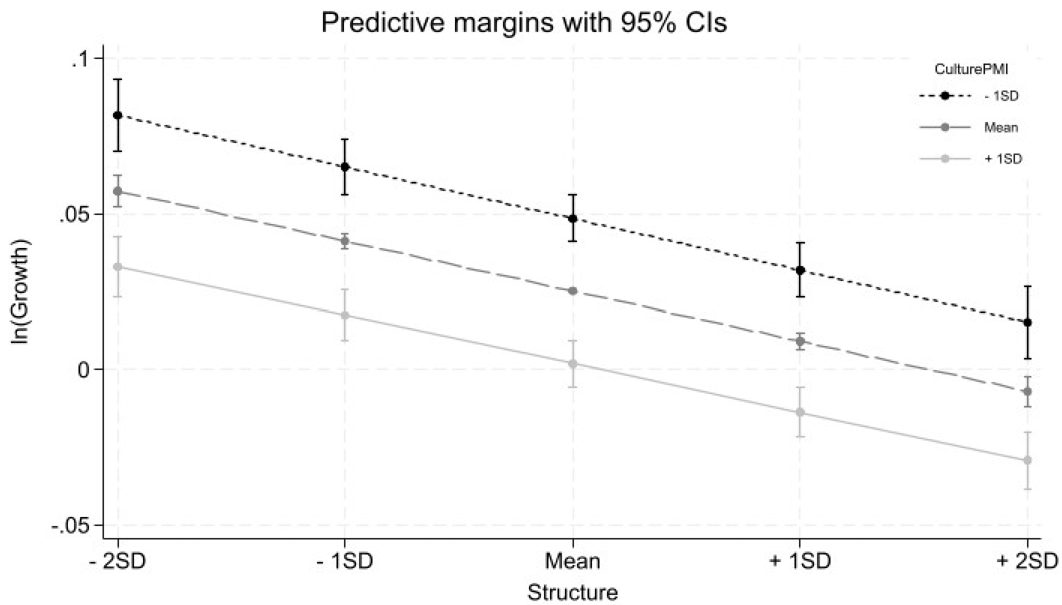


Figure 8: Predicted Growth Outcomes - CultureW2V Measure

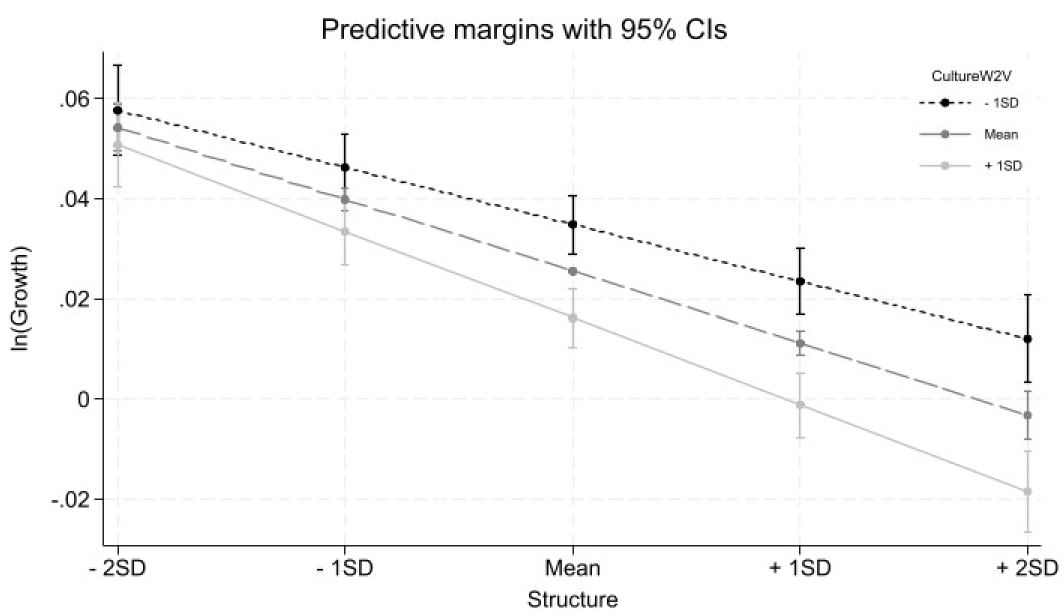


Figure 9: Predicted Engagement Outcomes - CulturePMI Measure

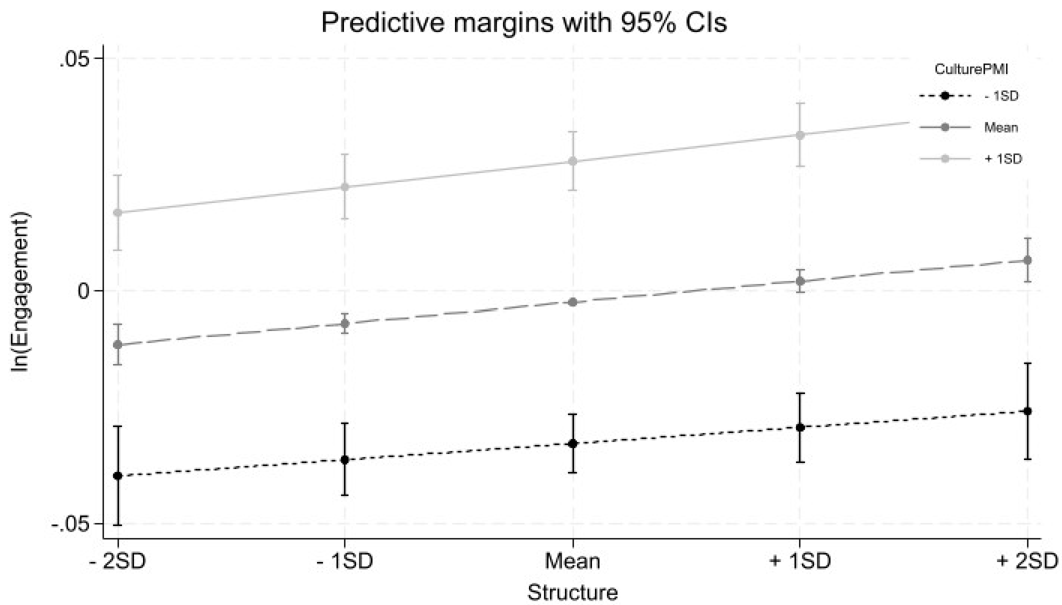
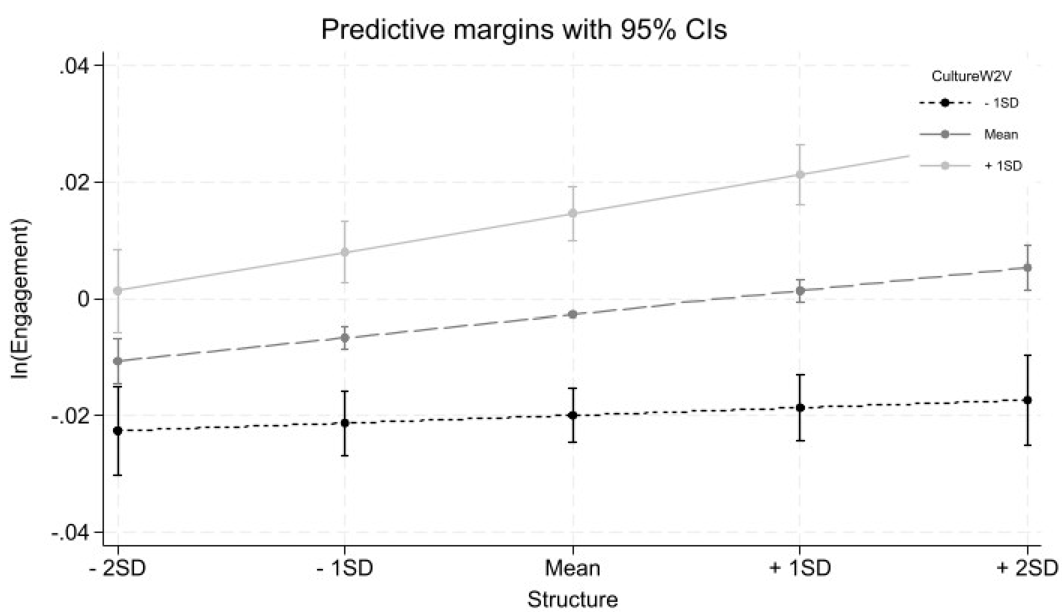


Figure 10: Predicted Engagement Outcomes - CultureW2V Measure



References

Adamic LA, Huberman BA (2000) Power-law distribution of the world wide web. *Science* 287(5461):2115–2115, publisher: American Association for the Advancement of Science.

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

Aral S, Walker D (2014) Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science* 60(6):1352–1370, publisher: INFORMS.

Baldassare M (1992) Suburban communities. *Annual Review of Sociology* 18(1):475–494, publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

Barabási AL (2009) Scale-free networks: a decade and beyond. *Science* 325(5939):412–413, publisher: American Association for the Advancement of Science.

Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512, publisher: American Association for the Advancement of Science.

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

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

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

Bhattacharjee S, Gopal RD, Lertwachara K, Marsden JR, 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, publisher: INFORMS.

Breiger RL (1974) The Duality of Persons and Groups. *Social Forces* 53(2):181–190, ISSN 0037-7732, URL `http://dx.doi.org/10.2307/2576011`, publisher: Oxford University Press.

Brynjolfsson E, Hu Y, 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, publisher: INFORMS.

Brynjolfsson E, Hu Y, Smith MD (2010) Research commentary—long tails vs. superstars: The effect of information technology on product variety and sales concentration patterns. *Information Systems Research* 21(4):736–747, publisher: INFORMS.

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

Caren N, Andrews KT, Lu T (2020) Contemporary social movements in a hybrid media environment. *Annual Review of Sociology* 46:443–465, publisher: Annual Reviews.

Carlson NA (2023) Differentiation in microenterprises. *Strategic Management Journal* 44(5):1141–1167, publisher: Wiley Online Library.

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

Carroll GR, Dobrev SD, Swaminathan A (2002) Organizational processes of resource partitioning. *Research in Organizational Behavior* 24:1–40, ISSN 01913085, URL `http://dx.doi.org/10.1016/S0191-3085(02)24002-2`.

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

Casciaro T, Piskorski MJ (2005) Power imbalance, mutual dependence, and constraint absorption: A closer look at resource dependence theory. *Administrative Science Quarterly* 50(2):167–199, publisher: SAGE Publications Sage CA: Los Angeles, CA.

Castells M (2015) *Networks of outrage and hope: Social movements in the Internet age* (Cambridge, UK: Polity).

Chen PL, Williams C, Agarwal R (2012) Growing pains: Pre-entry experience and the challenge of transition to incumbency. *Strategic Management Journal* 33(3):252–276, ISSN 1097-0266, URL `http://dx.doi.org/10.1002/smj.954`, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.954.

Dewan S, Ramaprasad J (2014) Social media, traditional media, and music sales. *Mis Quarterly* 38(1):101–122, publisher: JSTOR.

DiMaggio P, Hargittai E, Neuman WR, Robinson JP (2001) Social implications of the Internet. *Annual Review of Sociology* 27(1):307–336, publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

Dobrev SD, Kim TY, Carroll GR (2002) The Evolution of Organizational Niches: U.S. Automobile Manufacturers, 1885–1981. *Administrative Science Quarterly* 47(2):233–264, ISSN 0001-8392, 1930-3815, URL `http://dx.doi.org/10.2307/3094805`.

Elberse A, 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, Goodwin J (1994) Network analysis, culture, and the problem of agency. *American Journal of Sociology* 99(6):1411–1454, publisher: University of Chicago Press.

Evans JA, Aceves P (2016) Machine translation: Mining text for social theory. *Annual review of sociology* 42:21–50, publisher: Annual Reviews.

Faraj S, Jarvenpaa SL, 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, ISSN 0002-9602, 1537-5390, URL `http://dx.doi.org/10.1086/225993`.

Fisher G (2019) Online communities and firm advantages. *Academy of Management Review* 44(2):279–298, publisher: Academy of Management Briarcliff Manor, NY.

Fitzgerald B (2006) The transformation of open source software. *MIS Quarterly* 30(3):587–598, publisher: JSTOR.

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

Geroski PA (2005) Understanding the implications of empirical work on corporate growth rates. *Managerial and Decision Economics* 26(2):129–138, ISSN 1099-1468, URL `http://dx.doi.org/10.1002/mde.1207`, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mde.1207.

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

Guzman J, Li A (2023) Measuring founding strategy. *Management Science* 69(1):101–118, publisher: INFORMS.

Hamilton WL, Clark K, Leskovec J, Jurafsky D (2016) Inducing domain-specific sentiment lexicons from unlabeled corpora. *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*, volume 2016, 595 – 605 (NIH Public Access).

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

Hosanagar K, Fleder D, Lee D, Buja A (2014) Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. *Management Science* 60(4):805–823, publisher: INFORMS.

Hwang S, Foote JD (2021) Why do people participate in small online communities? *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–25, publisher: ACM New York, NY, USA.

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

Johnson SL, Safadi H, Faraj S (2015) The emergence of online community leadership. *Information Systems Research* 26(1):165–187, publisher: INFORMS.

Kozlowski AC, Taddy M, Evans JA (????) The Geometry of Culture: Analyzing Meaning through Word Embeddings. *American Sociological Review* 84(5):73.

Lifshitz-Assaf H (2018) Dismantling Knowledge Boundaries at NASA: The Critical Role of Professional Identity in Open Innovation. *Administrative Science Quarterly* 63(4):746–782, ISSN 0001-8392, URL `http://dx.doi.org/10.1177/0001839217747876`, publisher: SAGE Publications Inc.

Lin Z, Salehi N, Yao B, Chen Y, Bernstein M (2017) Better when it was smaller? community content and behavior after massive growth. *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 132–141, issue: 1.

Macdonald D (1953) A theory of mass culture. *Diogenes* 1(3):1–17, publisher: Sage Publications Sage CA: Thousand Oaks, CA.

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

Martin T (2017) community2vec: Vector representations of online communities encode semantic relationships. *Proceedings of the Second Workshop on NLP and Computational Social Science*, 27–31.

McKendrick DG, Hannan MT (2014) Oppositional identities and resource partitioning: Distillery ownership in Scotch whisky, 1826–2009. *Organization Science* 25(4):1272–1286, publisher: Informs.

McPherson JM, Rotolo T (1996) Testing a dynamic model of social composition: Diversity and change in voluntary groups. *American Sociological Review* 61(2):179–202, publisher: JSTOR.

McPherson JM, Smith-Lovin L (1987) Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Sociological Review* 52(3):370–379, publisher: JSTOR.

McPherson M (1983) An Ecology of Affiliation. *American Sociological Review* 48(4):519–532, ISSN 0003-1224, URL `http://dx.doi.org/10.2307/2117719`, publisher: [American Sociological Association, Sage Publications, Inc.].

Mickeler M, Khashabi P, Kleine M, Kretschmer T (2023) Knowledge seeking and anonymity in digital work settings. *Strategic Management Journal* 44(10):2113 – 2442, publisher: Wiley Online Library.

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

Neeley TB, Leonardi PM (2018) Enacting knowledge strategy through social media: Passable trust and the paradox of nonwork interactions. *Strategic Management Journal* 39(3):922–946, publisher: Wiley Online Library.

Negro G, Visentin F, Swaminathan A (2014) Resource partitioning and the organizational dynamics of "fringe banking". *American Sociological Review* 79(4):680–704, publisher: Sage Publications Sage CA: Los Angeles, CA.

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, publisher: SIAM.

Park Y, Bang Y, Ahn JH (2020) How does the mobile channel reshape the sales distribution in E-commerce? *Information Systems Research* 31(4):1164–1182, publisher: INFORMS.

Peli G, Nooteboom B (1999) Market partitioning and the geometry of the resource space. *American Journal of Sociology* 104(4):1132–1153, publisher: The University of Chicago Press.

Peterson A, Wu A (2021) Entrepreneurial learning and strategic foresight. *Strategic Management Journal* 42(13):2357–2388, publisher: Wiley Online Library.

Peterson RA, Anand N (2004) The production of culture perspective. *Annu. Rev. Sociol.* 30:311–334, publisher: Annual Reviews.

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

Podolny JM, Stuart TE, Hannan MT (1996) Networks, Knowledge, and Niches: Competition in the Worldwide Semiconductor Industry, 1984-1991. *American Journal of Sociology* 102(3):659–689, ISSN 0002-9602, URL `https://www.jstor.org/stable/2782460`, publisher: University of Chicago Press.

Porter CE, Donthu N (2008) Cultivating trust and harvesting value in virtual communities. *Management Science* 54(1):113–128, publisher: INFORMS.

Ren Y, Harper FM, Drenner S, Terveen L, Kiesler S, Riedl J, Kraut RE (2012) Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *MIS Quarterly* 36(3):841–864, publisher: JSTOR.

Ruef M (2000) The Emergence of Organizational Forms: A Community Ecology Approach. *American Journal of Sociology* 106(3):658–714, ISSN 0002-9602, 1537-5390, URL `http://dx.doi.org/10.1086/318963`.

Salganik MJ, Dodds PS, Watts DJ (2006) Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* 311(5762):854–856, ISSN 0036-8075, 1095-9203, URL `http://dx.doi.org/10.1126/science.1121066`.

Shils E (1962) The theory of mass society: Prefatory remarks. *Diogenes* 10(39):45–66, publisher: Sage Publications Sage CA: Thousand Oaks, CA.

Sine WD, Haveman HA, Tolbert PS (2005) Risky business? Entrepreneurship in the new independent-power sector. *Administrative Science Quarterly* 50(2):200–232, publisher: SAGE Publications Sage CA: Los Angeles, CA.

Srivastava SB, Goldberg A (2017) Language as a window into culture. *California Management Review* 60(1):56–69, publisher: SAGE Publications Sage CA: Los Angeles, CA.

Srivastava SB, Goldberg A, Manian VG, Potts C (2018) Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Science* 64(3):1348–1364, publisher: INFORMS.

Stewart D (2005) Social status in an open-source community. *American Sociological Review* 70(5):823–842, publisher: Sage Publications Sage CA: Los Angeles, CA.

Swaminathan A (1995) The proliferation of specialist organizations in the American wine industry, 1941-1990. *Administrative science quarterly* 40(4):653–680, publisher: JSTOR.

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

Tan TF, Netessine S, 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, publisher: INFORMS.

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

TeBlunthuis N, Kiene C, Brown I, Levi L, McGinnis N, Hill BM (2022) No community can do everything: why people participate in similar online communities. *Proceedings of the ACM on Human-Computer Interaction* 6(CSCW1):1–25, publisher: ACM New York, NY, USA.

Tortoriello M, Reagans R, 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, publisher: INFORMS.

Van Alstyne M, Brynjolfsson E (2005) Global village or cyber-balkans? Modeling and measuring the integration of electronic communities. *Management Science* 51(6):851–868, publisher: INFORMS.

Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(11):2579 – 2605.

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

Wang X, Butler BS, Ren Y (2013) The Impact of Membership Overlap on Growth: An Ecological Competition View of Online Groups. *Organization Science* 24(2):414–431, ISSN 1047-7039, 1526-5455, URL `http://dx.doi.org/10.1287/orsc.1120.0756`.

Wasko MM, Faraj S (2005) Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly* 29(1):35–57, publisher: JSTOR.

Wellman B (2001) Computer networks as social networks. *Science* 293(5537):2031–2034, publisher: American Association for the Advancement of Science.

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

Zald MN, Ash R (1966) Social movement organizations: Growth, decay and change. *Social Forces* 44(3):327–341, publisher: The University of North Carolina Press.

Zentner A, Smith M, Kaya C (2013) How video rental patterns change as consumers move online. *Management Science* 59(11):2622–2634, publisher: INFORMS.

Zhang J, Hamilton W, Danescu-Niculescu-Mizil C, Jurafsky D, Leskovec J (2017) Community identity and user engagement in a multi-community landscape. *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 377–386, issue: 1.

Zhang L (2018) Intellectual property strategy and the long tail: Evidence from the recorded music industry. *Management Science* 64(1):24–42, publisher: INFORMS.

Zhu DH, Chen G (2015) Narcissism, director selection, and risk-taking spending. *Strategic Management Journal* 36(13):2075–2098, publisher: Wiley Online Library.

Zhu H, Chen J, Matthews T, Pal A, Badenes H, Kraut RE (2014) Selecting an effective niche: an ecological view of the success of online communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 301–310.

Zuckerman EW, Kim TY (2003) The critical trade-off: identity assignment and box-office success in the feature film industry. *Industrial and Corporate Change* 12(1):27–67, publisher: Oxford University Press.

Appendix A: Sample Filtering Criteria

In creating the sample, I used a number of filtering criteria—at the level of the comment, user, and subreddit. Many of these filters follow other studies of Reddit and other similar online communities (e.g., Wang et al. 2013, Zhang et al. 2017, Waller and Anderson 2021, Teblunthuis and Hill 2022). First, in terms of comments, on Reddit, some comments are removed by moderators or by the user. In addition, users can delete their account or be banned from Reddit, meaning that some comments lose the information on their associated author. Comments without text or author data would not allow for calculation of key independent variables around structure and culture, and thus I chose to remove these comments from the sample.

Second, in terms of users, this is a study of user time and attention as the key resource for online communities on Reddit. As such, I am interested in examining attention allocation within the site, rather than between Reddit and other external sites. I therefore wanted to make sure that I was capturing within-platform attention. To do so, I followed other studies in limiting the analysis to users who made a certain number of comments, above a minimum threshold (Hamilton et al. 2017). Specifically, I removed users who posted fewer than 10 comments in their lifetime on Reddit (Waller and Anderson 2021). Another reason for doing this was that I am interested in competition for attention on Reddit, but only attention that represents a potential significant resource, rather than small pieces of attention by very transitory users.

Third, at the subreddit level, I filtered subreddits based on a number of criteria. First, I removed subreddits that existed for only a single month, doing so in order to remove junk and transient subreddits. As another measure to remove junk subreddits and subreddits dominated by bots, I removed subreddits that never grew beyond a single user in their entire lifetime within the study period, as these were most likely to be bot-dominated subreddits. Additionally, I removed subreddits where the average number of comments per user was above 2,500 per month—this is because this is a reasonable upper-level threshold for real humans to comment per month (assuming the most intense Reddit users are using Reddit about 10 hours per day, 6 days per week, and making roughly 10 comments/hour, this equates to around 2,500 comments per month). Subreddits above this threshold are very likely to be dominated by bots. Inspection of subreddits corroborated this estimate, with the most engaged, primarily human-populated subreddits garnering roughly 2,500 comments per user per month. Above this threshold, subreddits are either fully spam (e.g. users spamming the same comment over and over) or completely dominated by bots (such as a subreddit in which a bot posted match recaps from DOTA2). I eliminated these junk and bot-centric subreddits, as they were not the focus of the analyses. Instead, I wanted to focus on subreddits consisting mostly of human users. There were 27 of these subreddits. I also ran models with these subreddits included, and findings were consistent. Note that I do not take any steps to remove bots beyond these subreddits, as bots are part of the ecosystem of Reddit (and the product of human attention) and contribute content that can help shape communities. I only wanted to remove communities in which there was no direct human participation.

As additional filters, I included only subreddits that met a minimum threshold of activity—specifically those that attracted an average of at least 100 comments per month. The reasoning was to remove subreddits that did not draw enough activity to enter into the competitive ecology in a real way (Wang et al. 2013). In addition, I considered a subreddit inactive in a given month if it had fewer than 10 comments in that month. I did this in order to limit the sample of subreddits each month to those with enough text data to reasonably calculate culture measures. Additionally, I ran a model that tested the language of each subreddit, and included only those in English (Wang et al. 2013, Waller and Anderson 2021), doing so because the culture measures may operate differently for non-English text. Even after these filters, there were 3 subreddit-months in which I could not calculate a culture score. This is because the measures for culture included only words appearing at least 5 times on Reddit in a particular month. These subreddits had no words that reached this threshold, and so I could not calculate a culture score. As such, I did not include them in the dataset.

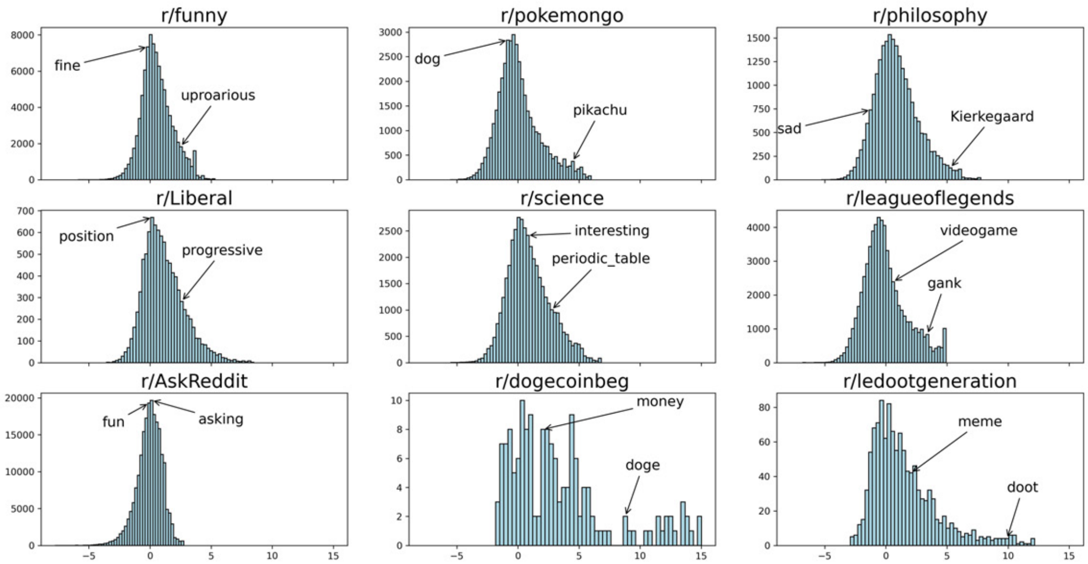
Appendix B: Additional Details on CulturePMI Measure

In this section, I offer additional detail into the CulturePMI measure. Underlying the measure for each subreddit, is a measure of how distinct each word is to the particular subreddit (Zhang et al. 2017). When a subreddit contains a larger proportion of words that are distinct to that subreddit relative to other subreddits, then the subreddit will be more distinct overall. Stated differently, subreddits where discussion uses more localized language are more distinct.

In order to give some more insight into what this means, I picked a sample of 9 subreddits. The sample subreddits are: r/funny, r/pokemongo, r/philosophy, r/Liberal, r/science, r/leagueoflegends, r/AskReddit, r/dogecoinbeg, and r/ledootgeneration. For each subreddit, I plotted the distribution of word distinctiveness scores for each subreddit (these are the scores which are averaged to get the overall culture score for the subreddit). A higher score indicates that a word is used more frequently in the focal subreddit than in other subreddits, indicating the word is distinct to that subreddit context. A score close to 0 indicates that the word is used in relatively similar frequencies in the focal subreddit and in other subreddits. Finally, a score below 0 indicates that the word is used relatively less frequently in the focal subreddit relative to other subreddits. Many subreddits have a distribution centered around zero, but some are more skewed. For example, r/AskReddit, which is the largest subreddit, has a right-skewed distribution, indicating a lack of many distinct words, underlying its low overall culture score. On the opposite end of the spectrum, r/ledootgeneration has a longer tail of distinct words, associated with its high culture score. Similarly, r/dogecoinbeg, a very small community, has many distinct words and thus has a high culture score.

On the word score distributions for each subreddit, I identify two words. For most subreddits, I picked a word that exemplified the subreddit and thus was likely to have a high score, as well as a word that was likely to not be so distinct to the particular subreddit, and thus would have a lower score. For the r/funny subreddit, a distinct word is "uproarious," whereas a less distinct word is "fine." For the r/pokemongo subreddit, focused on the video game Pokemon Go, a distinct word is the pokemon "pikachu" whereas a less distinct word is the generic animal "dog." For the r/philosophy subreddit, a distinct word is the famous philosopher "Kierkegaard," whereas a more generic feeling "sad" is less not distinct. For the r/Liberal subreddit, "progressive" is a relatively distinct word, whereas "position" is not. For the subreddit r/science, "periodic table" is a relatively distinct word, whereas "interesting" is not. For the r/leagueoflegends subreddit, focused on the popular video game, "gank" is a distinct word whereas "videogame" is not. For the r/AskReddit subreddit, the largest subreddit which is a generic ask and answer forum, I chose two words "asking" and "fun" but both are relatively non-distinct. This subreddit does not have a large proportion of distinct words, as it contains mostly generic discussion. For the r/dogecoinbeg subreddit, focused around people asking others to give them coins in the cryptocurrency Dogecoin, "doge" is a very distinct word, whereas "money" is not. For the r/ledootgeneration subreddit, a very distinct subreddit focused around a particular meme involving a skeleton playing "doot doot" notes on a trumpet, the word "doot" is quite distinct, whereas "meme" is less so. These graphs give greater insight into what the CulturePMI measure is actually capturing.

Figure 11: CulturePMI Examples



Appendix C: Additional Details on CultureW2V Measure

In this section, I offer additional detail on the word embedding model used to create the CultureW2V scores. I leveraged the skip-gram architecture of the Word2Vec algorithm, trained with negative sampling procedure (5 negative samples), which has been shown to have superior performance on large datasets (Mikolov et al. 2013). Skip-gram uses a shallow, two-layered neural network architecture, which enables efficient training even on massive data sets. The model reads in a corpus line-by-line, utilizing a sliding window of k words (I used k=5 for this study), and predicts the context surrounding each word—meaning specifically that it predicts the words, within window k, surrounding the focal input word (Mikolov et al. 2013, Kozlowski et al. 2019). Additionally, I trained models over 5 epochs (iterations), to increase quality of word representations.

To train skip-gram models on Reddit data, I used the python library Gensim. For each month, I transformed the Reddit data into a text file in which each comment represented a “document” in the monthly corpus. Skip-gram models were thus trained on within-comment context windows, to generate vectors based on words actually co-occurring in the data. This resulted in a word-embedding model for each month of data.

Beyond these details, I include here, in Figure 12, the same t-SNE plot included in the main text of subreddits based on their vectors underlying the CultureW2V score, but here I include text of the names of a random sample of subreddits, to show the validity of the similarity calculations. I ran a K-Means clustering algorithm, and then randomly sampled 1.5 percent of subreddits within each cluster (percentage chosen to balance coverage and text visibility). Inspecting subreddit names reveals clusters of topically-similar subreddits.

In addition, I include Figure 13 which displays 10 sample words and their most relevant matches based on cosine similarity from the word embedding model trained on December 2016 data. This gives a sense of what word similarity means in the word2vec model.

Finally, I include Figure 14, which displays a t-SNE plot based on embeddings for words. I include here only the most frequent 10,000 words (out of 500,000 total words), from December 2016, excluding the top 100 most frequent words. The size of each word's dot represents the frequency of that word. I then ran a K-means clustering algorithm, and included 10 percent of words from each cluster, to give a sense of the word similarities. This plot shows the underlying relationship between words, which in turn underlie the subreddit vectors that go into the CultureW2V score. Similarly, here inspection reveals clusters of semantically-similar words.

Figure 12: CultureW2V Plot

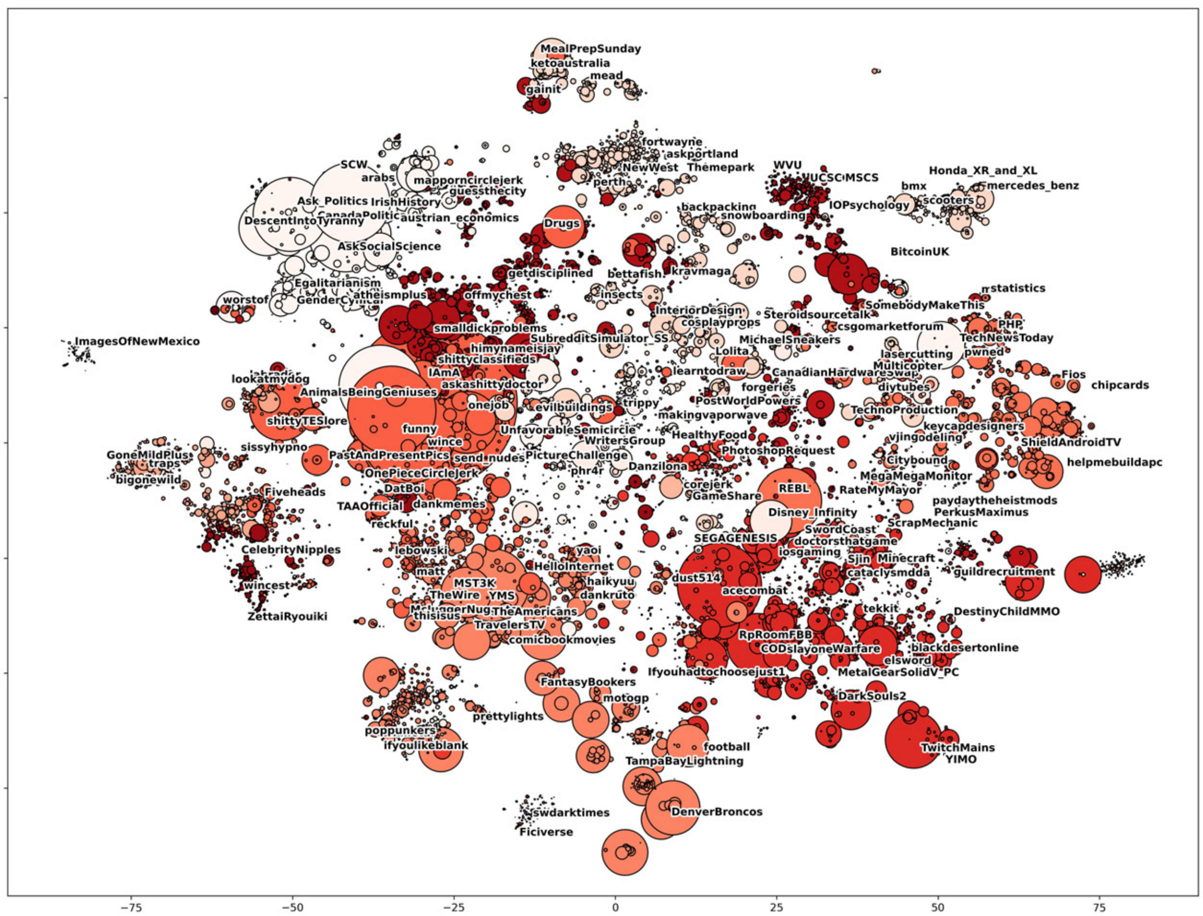
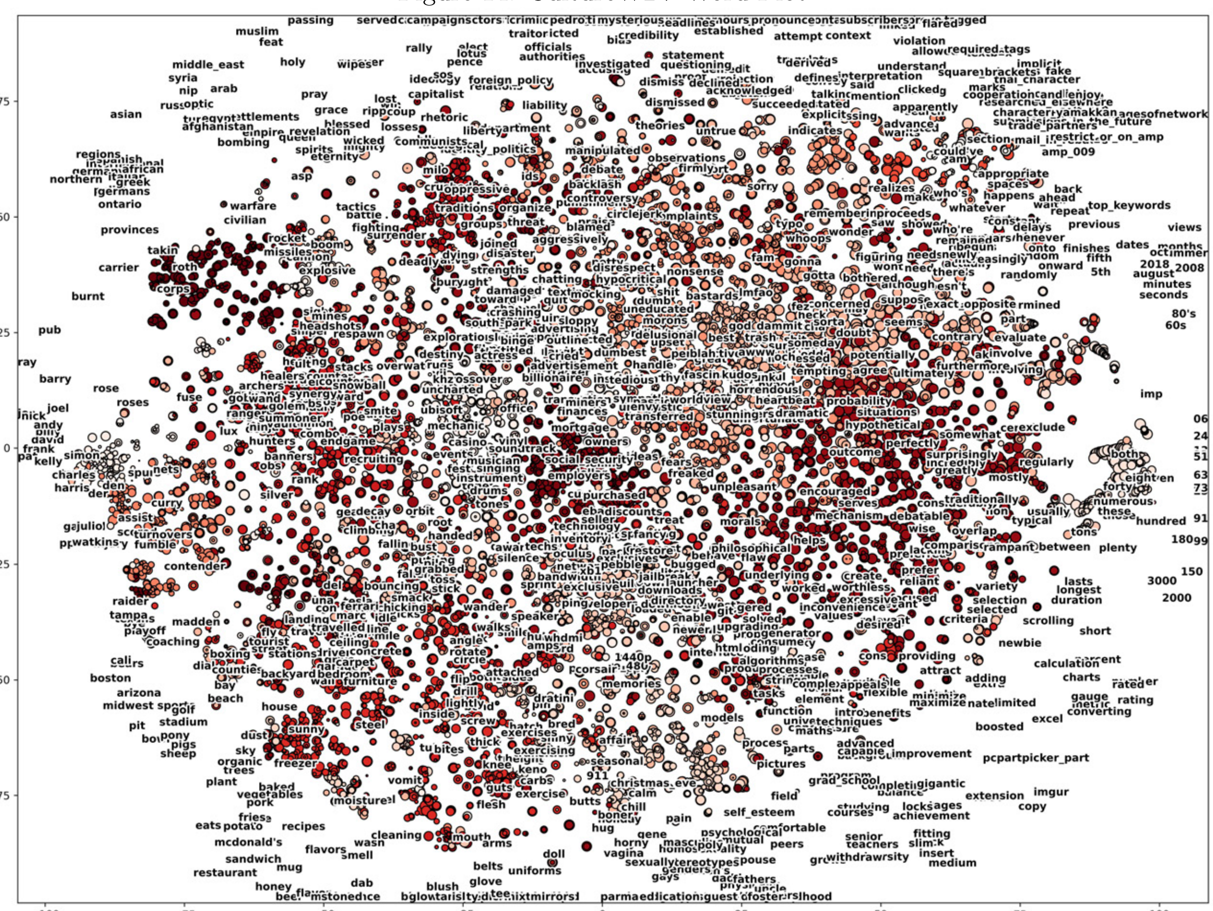


Figure 13: CultureW2V Words

bank	disease	elephant	fad	ghost	louis_vuitton	nintendo	outerspace	superman	yankees
banks	diseases	elephants	fads	ghosts	kate_spade	nintendo's	outer_space	superman's	orioles
banking	incurable	rhinoceros	craze	ghostly	michael_kors	nintendo_consoles	alien_spaceship	batman	dodgers
fdic_insured	illnesses	giraffe	trendy	25_netherworld	handbags	3ds_successor	spaceship	darkseid	braves
banker	inflammatory_bowel	tiger	hipsters	haunted	zegna	nintendo_handhelds	interplanetary_travel	deathstroke	blue_jays
chequing	curable	elephant_trunk	trend	fuzzy_skull	tommy_hilfiger	consoles	alien_spacecraft	dcau	cubs
debited	illness	animal	hipster	fuzzy_vamp	louboutin	wiiu's	terraform_mars	martian_manhunter	padres
financial_institution	communicable_diseases	gag_gifts	fad_diet	haunting	rolex_watches	virtual_console	aliens	lex_author	astros
debit_card	prion_diseases	stuffed_animal	trendiness	spooky_ghost	handbag	sony_and_microsoft	rocketship	wonderwoman	mariners
icici_bank	untreatable	saver_tooth	uncool	shuppet	hilfiger	wii's	moonbase	batman's	josh_donaldson
wire_transfers	chronic_diseases	taxidermied	hipsterism	meliora	lacoste	nintendon't	spaceships	lois_and_clark	cashman

Figure 14: CultureW2V Word Plot



Appendix D: Results from Alternative Measure of Structure

Table 7: Growth Models with Alternative Structure Measure

	(1)	(2)	(3)	(4)	(5)
Probability of Being Active	3.196***	3.152***	3.196***	2.907***	3.347***
	(0.346)	(0.345)	(0.345)	(0.345)	(0.345)
Num Users	-0.248***	-0.254***	-0.244***	-0.246***	-0.243***
	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)
Comments per User	0.091***	0.087***	0.094***	0.087***	0.094***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Comment Length	-0.043***	-0.050***	-0.044***	-0.037***	-0.038***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Default	0.192***	0.178***	0.174***	0.222***	0.181***
	(0.019)	(0.018)	(0.018)	(0.021)	(0.019)
Age (Months)	0.005***	0.005***	0.005***	0.004***	0.005***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mod Activity	-0.005	-0.004	-0.005	-0.003	0.000
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Structure		-0.033***	-0.033***	-0.038***	-0.338***
		(0.003)	(0.003)	(0.003)	(0.022)
CulturePMI		-0.046***		0.104***	
		(0.007)		(0.011)	
CultureW2V			-0.020***		0.086***
			(0.006)		(0.010)
Structure × CulturePMI				-0.125***	
				(0.007)	
Structure × CultureW2V					-0.093***
					(0.006)
Constant	1.089***	1.176***	1.031***	1.099***	1.355***
	(0.020)	(0.024)	(0.027)	(0.025)	(0.036)
Time FE	Yes	Yes	Yes	Yes	Yes
Subreddit FE	Yes	Yes	Yes	Yes	Yes
Observations	440,235	440,235	440,235	440,235	440,235

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 8: Growth Margins with Alternative Structure Measure

	CulturePMI	CultureW2V
-2SD	0.0515***	0.0477***
	(0.009)	(0.008)
-1SD	0.00218	0.0112
	(0.008)	(0.007)
Mean	-0.0471***	-0.0252***
	(0.008)	(0.006)
+1SD	-0.0964***	-0.0617***
	(0.008)	(0.007)
+2SD	-0.146***	-0.0981***
	(0.009)	(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 9: Engagement Models with Alternative Structure Measure

	(1)	(2)	(3)	(4)	(5)
Probability of Being Active	5.265***	5.314***	5.251***	5.211***	5.238***
	(0.229)	(0.228)	(0.228)	(0.228)	(0.228)
Num Users	0.056***	0.068***	0.056***	0.071***	0.056***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Comments per User	-0.434***	-0.427***	-0.436***	-0.427***	-0.436***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Comment Length	-0.023***	-0.013**	-0.020***	-0.007	-0.020***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Default	-0.086***	-0.084***	-0.080***	-0.066***	-0.080***
	(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	0.001	-0.000	0.000	0.000	0.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Structure		0.017***	0.017***	0.015***	0.042**
		(0.002)	(0.002)	(0.002)	(0.016)
CulturePMI		0.061***		0.123***	
		(0.006)		(0.009)	
CultureW2V			0.036***		0.028***
			(0.005)		(0.007)
Structure × CulturePMI				-0.052***	
				(0.005)	
Structure × CultureW2V					0.008
					(0.005)
Constant	0.221***	0.106***	0.323***	0.074***	0.296***
	(0.014)	(0.021)	(0.018)	(0.021)	(0.026)
Time FE	Yes	Yes	Yes	Yes	Yes
Subreddit FE	Yes	Yes	Yes	Yes	Yes
Observations	440,235	440,235	440,235	440,235	440,235

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 10: Engagement Margins with Alternative Structure Measure

	CulturePMI	CultureW2V
-2SD	0.102***	0.0308***
	(0.008)	(0.006)
-1SD	0.0811***	0.0338***
	(0.007)	(0.005)
Mean	0.0605***	0.0368***
	(0.006)	(0.005)
+1SD	0.0399***	0.0398***
	(0.007)	(0.005)
+2SD	0.0193*	0.0429***
	(0.008)	(0.006)

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

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

Appendix E: Results from Alternative Measure of Engagement

Table 11: Alternate Engagement Models

	(1)	(2)	(3)	(4)	(5)
Probability of Being Active	-0.905***	-0.850***	-0.912***	-0.849***	-0.877***
	(0.203)	(0.201)	(0.203)	(0.201)	(0.203)
Num Users	0.001	0.018***	0.003*	0.018***	0.003*
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Comments per User	-0.003	0.007	-0.004	0.007	-0.004
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Comment Length	-0.683***	-0.670***	-0.681***	-0.670***	-0.682***
	(0.006)	(0.007)	(0.006)	(0.007)	(0.006)
Default	0.022	0.013	0.021	0.013	0.019
	(0.014)	(0.013)	(0.014)	(0.013)	(0.014)
Age (Months)	-0.000	-0.000	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mod Activity	0.009***	0.007***	0.008***	0.007***	0.008***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Structure		0.007***	0.007***	0.006***	0.054***
		(0.001)	(0.001)	(0.001)	(0.009)
CulturePMI		0.071***		0.068***	
		(0.009)		(0.017)	
CultureW2V			0.023***		-0.055***
			(0.006)		(0.014)
Structure × CulturePMI				0.001	
				(0.003)	
Structure × CultureW2V					0.015***
					(0.003)
Constant	2.335***	2.171***	2.369***	2.172***	2.132***
	(0.022)	(0.034)	(0.022)	(0.034)	(0.051)
Time FE	Yes	Yes	Yes	Yes	Yes
Subreddit FE	Yes	Yes	Yes	Yes	Yes
Observations	440,235	440,235	440,235	440,235	440,235

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 12: Alternate Engagement Margins

	CulturePMI	CultureW2V
-2SD	0.0702***	0.00112
	(0.010)	(0.007)
-1SD	0.0706***	0.0111
	(0.009)	(0.006)
Mean	0.0710***	0.0210***
	(0.009)	(0.006)
+1SD	0.0713***	0.0309***
	(0.009)	(0.007)
+2SD	0.0717***	0.0409***
	(0.010)	(0.007)

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

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