Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature Date

_____ _____

Pranay Jinna

Essays on Membership and Participation in Online Communities

By

Pranay Jinna
Doctor of Philosophy
Business

_____
Anand Swaminathan, Ph.D.
Advisor

_____
Anandhi Bharadwaj, Ph.D.
Advisor

_____
Benn Konsynski, Ph.D
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____
Date

Essays on Membership and Participation in Online Communities

By

Pranay Jinna
B.E., Osmania University, 2004
MEM, Duke University, 2008

Advisors
Anand Swaminathan, Ph.D.
Anandhi Bharadwaj, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Business
2019

Abstract

Essays on Membership and Participation in Online Communities

By Pranay Jinna

My dissertation consists of two essays focusing on membership and participation in online communities. In my first essay titled “Attraction, Participation and Retention in Online Communities: An Ecological Overview”, I draw from theories of user participation and organizational ecology to explain how overlapping membership in online communities affects attraction, participation and long-term retention of members in these communities. I find that high membership overlap density has a negative effect on new member attraction but a surprising positive effect on long term retention of existing members. Sharing members with other communities increases long term retention of members. Further analysis at the individual level shows us that members in multiple communities are more likely to stay longer on the platform and hence also stay longer in each of the communities. I find that as multiple membership increases, members are more likely to increase their overall engagement on the platform but their engagement per community decreases. In my second paper titled “Bias in Online Reviews: Variety and Atypicality in Online Gaming”, I study self-selection biases in online reviews while accounting for review propensity of different segments of individuals. Segmenting individuals based on their variety seeking and atypicality seeking preferences, I find that poly-mixers - individuals with high variety seeking and atypicality seeking tendencies are less likely to review products but more likely to give lower ratings compared to other individuals. I find that an individual’s social network, the number of products owned and their product usage significantly impact their propensity to review and rate products.

Essays on Membership and Participation in Online Communities

By

Pranay Jinna
B.E., Osmania University, 2004
MEM, Duke University, 2008

Advisors
Anand Swaminathan, Ph.D.
Anandhi Bharadwaj, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Business
2019

Table of Contents

Essay 1: Attraction, Participation and Retention in Online Communities: An Ecological Perspective

Essay 2: Bias in Online Reviews: Variety and Atypicality in Online Gaming

Attraction, Participation and Retention in Online Communities: An Ecological Perspective

Abstract

In this paper, I draw from theories of user participation and organizational ecology to explain how overlapping membership in online communities affects attraction, participation and long term retention of members in these communities. I conduct analysis, at the community level and at the individual level, on 797 communities over a 36 month period comprising over 500,000 individuals. I find that high membership overlap density has a negative effect on new member attraction but a surprising positive effect on long term retention of existing members. Sharing members with other communities increases long term retention of members. Further analysis at the individual level shows us that members in multiple communities are more likely to stay longer on the platform and hence also stay longer in each of the communities. I also take advantage of an exogenous event, where the platform owners permitted members to create and moderate communities on any topic leading to the creation of a large number of new communities, to study the causal effect of multiple membership on member engagement. I find that as multiple membership increases, members are more likely to increase their overall engagement on the platform but their engagement per community decreases. These findings have important implications from a theoretical and managerial perspective for understanding long term member engagement in online communities.

1. Introduction

The widespread adoption of the Internet has allowed homophilous groups of people to come together to discuss ideas, opinions, news and other articles of shared interest on online forums (Sproull and Arriaga 2007). These online communities evolve organically and establish their own processes, structures and norms and they change over time with members entering and exiting the community. Membership in most of these communities is voluntary, and individuals, who are often anonymous, have the option to join and quit a community at any point of time (Moon and Sproull 2008). The success of a community depends on its retention of its existing members and/or its ability to attract new members. Researchers studying the sustainability of online communities have examined factors such as individual motivation for participation (Wasko and Faraj 2000, Wasko and Faraj 2005), individual and group commitment (Bateman et al. 2011), leadership emergence in communities (Johnson et al. 2015), network exchange patterns (Faraj and Johnson 2011) and community level factors such as size and communication activity (Butler 2001). This line of research has focused mainly on the characteristics of individual members within a community and how an individuals' participation in a community and commitment towards that community help to create and build a thriving online community and how a community's norms and structure affect an individual's continued participation in that community.

A secondary line of research has considered community evolution from an ecological perspective and studies how a community's membership overlap on a platform, affects the community's sustainability (Wang et al. 2013, Zhu et al. 2014, Butler and Wang 2012). Wang et al. (2013) consider community evolution from the ecological view and study how competition defined by a community's membership overlap affects the growth of membership of the focal community. They study a subset of communities on Usenet between 1999 and 2005 and find that communities with higher member overlap experience lower monthly growth. Zhu et al. (2014) analyze 5,673 Wikia communities to study the effect of membership

overlap on survival of online communities and find that communities with higher membership overlap survive longer.

To reconcile these orthogonal findings, I propose a synthesis of two lines of research and study community sustainability both from the theory of individual participation and from the ecology of community membership. Prior researchers have focused mainly on how commitment to the community and active member engagement within a community play a crucial role in sustaining online groups (Bateman et al. 2011, Butler et al. 2007). In research on online communities, studying member participation across multiple communities has been less emphasized. In this study, I examine how individuals' multiple community membership affects their long term participation on the platform and also in their communities. I find that, controlling for all factors including overall participation, members who participate in multiple communities stay on the platform for a longer duration compared to members who participate in fewer communities.

Member overlap density, a community level construct to measure membership overlap of a community, can be decomposed into an individual level construct based on the extent of participation by individuals across communities. The multiple membership of a member in a given month is the total number of communities that the member belongs to in that month. A community's member overlap density is equal to the average of all its members' multiple memberships. On average, communities that have high member overlap density have members who belong to more communities compared to communities that have low member overlap density. Further, members who belong to many communities engage in communities differently compared to members who are in fewer communities. Studying community sustainability from the perspective of multiple membership can explain the paradox as to why communities that grow at a slower pace survive longer.

I distinguish the effect of multiple membership on long term retention in the community from the effect of short term retention and engagement in the community, and explain how this affects the growth

of communities. I find that communities with high overlap density attract fewer new members but have a higher retention of their existing members compared to communities with low overlap density. I also find that members who have high multiple membership are more likely to stay on the platform and hence within a community. This can explain why communities with high member overlap density exhibit slower growth but survive longer. Although these communities attract fewer new members, members of these communities are likely to stay longer on the platform and hence, on average, also likely to stay longer in their communities. I also find that as members increase their multiple membership, they are more likely to increase their overall engagement on the platform but decrease their engagement per community.

The rest of the paper is organized as follows: In the next section, I review the literature on online communities studied from both the ecological perspective and the individual participation perspective and propose a set of hypotheses. In Sections 3 and 4, I describe my data, variables and the regressions. In Section 5, I report the results of my analysis. In section 6, I discuss the theoretical and practical implications and conclude.

2. Theory and Hypotheses

Research on online communities examining factors affecting sustainability of communities can be broadly classified into two lines. First, most of the research has considered online communities from the perspective of the individual and researchers have focused on how individual motivation and group engagement within a community are important in building thriving communities (Preece and Ghozati 2001, Wasko and Faraj 2000, Wasko and Faraj 2005, Bateman et al. 2011).A second, more nascent line of research has studied community sustainability from the perspective of ecological evolution and how competition faced by a community defined by its membership overlap affects its growth and survival (Wang et al. 2013, Zhu et al. 2014). In this paper, I propose a synthesis of these two lines of research to study how community membership overlap and individual participation affect community sustainability.

2.1 Ecological View of Sustainability of Online Communities

The ecological view of online communities considers competition among communities for resources as analogous to competition among organizations for resources and draws from ecological theories to understand the growth of organizations. Although online groups differ from traditional organizations along several dimensions, prior researchers studying online communities have argued that just as organizations compete for a variety of resources such as physical assets, natural resources, labor and financial capital to sustainably carry out their operations, online communities compete for individual members for growth. The most important resource for an online community is its members and hence ecological theories of competition for organizations are useful and relevant to study the evolution and sustainability of online communities (Wang et al. 2013).

Direct competition between organizations is often represented through the overlap in their niches. An organization's niche is defined as an "n-dimensional" resource space where it obtains resources for survival and growth (MacArthur and Levins 1967). The extent of niche overlap between two organizations defines the extent to which organizations compete against each other for the same resources. In prior research, the niche width of an organizational form has been calculated based on the context of the industry. Researchers have studied niches of restaurants in terms of the profile of service offered (Freeman and Hannan 1983), of semiconductor companies as positions in a technology space (Podolny et al. 1996), of investment banks by the distributions of activities over industries (Park and Podolny 2000), and of automobile manufacturers by the ranges of engine sizes that they produced (Dobrev et al. 2001; Dobrev et al. 2003).

Organizational ecology researchers use overlap density - number of other organizations in a population that share a resource niche with the focal firm, to characterize an organizations' competitive environment (Baum and Singh 1994). Overlap density reflects the level of competition for resources that a focal firm needs for its survival and growth (Baum and Singh 1994). For online communities, the essential

resources include members' time and effort and the ability to attract new members and retain existing members. Hence, researchers studying online communities from the ecological view use member overlap density to define the niche width of a community and to characterize an online community's competitive environment (Wang et al. 2013). Member overlap density for a community is defined as the sum of number of shared members between the focal community and every other community with whom it shares at least one member, weighted by the number of members in the focal community (Wang et al. 2013).

During the early growth phase of a platform which hosts online communities, there is a constant stream of new members joining the platform. These members have the option to join and participate in one or more communities from among the multiple communities that are a part of the platform. Communities with broader niches compete with many other communities for potential new members as these new members have a larger list of similar communities that they can join (Popielarz and McPherson 1995). Communities with a more focused niche are more likely to attract potential new members as they do not have many competitor communities. Therefore, we expect a negative effect of member overlap density on new member attraction.

Hypothesis 1 (H1): *Member overlap density is negatively associated with new member attraction in an online community.*

Members' expectations about the community evolve as they join and participate in a community. Members decide to continue to participate in the community depending on the direction of the discussions and response of other community members (Joyce and Kraut 2006). Changes in topic consistency within a community affect user participation and retention (Butler et al. 2014). If there is a change in the direction of discussion or response of new members, then members can choose to move to other communities. The presence of other similar communities might play a strong role in whether members decide to continue in the community or leave and join other communities. A community with a narrow niche faces lower competition; its members have fewer options to choose from and hence would continue to participate in the

community even if there is a change in topic consistency. A community with a broad niche focusses on a variety of topics and has multiple other communities competing with it. Members of communities with a broad niche have more options to choose from if they wish to switch. Popierlaz and McPherson (1995) study membership in voluntary organizations and find that voluntary groups lose fastest those members who are subject to competition from other groups. Hence, we would expect the communities having broad niches to have a lower member retention rate compared to that of communities having narrower niches.

Hypothesis 2 (H2): *Member overlap density is negatively associated with member retention in an online community.*

In online communities, if older members communicate only among themselves, lurkers who observe the community before participating would find that new members are not really welcomed in the community. In Open Source Software, project age has been found to have a positive effect on user attraction but no significant effect on developer attraction (Chengalur-Smith et al. 2010). In many communities on USENET, new members were told to read the manual before participating in the community. Newer members would feel ignored and would likely move to other communities who would be more welcoming or exit the platform altogether. As communities age, they would be less welcoming of new members. We would expect that as communities get older they are less likely to attract new members.

Hypothesis 3 (H3): *Community Age is negatively associated with member attraction in an online community.*

As a community get older, members of the community are more likely to coalesce around defined topics of interest. The norms and processes of older communities are more established and members are more likely to adhere to these norms. Over time, with experience, community moderators learn to regulate communities effectively. They would only allow discussion topics which would be of interest to its members and would better at mitigating or controlling trolling. Older members who have participated in a community for long periods of time would also create personal bonds with other members and hence would

be less likely to leave the community. Further, older communities are also more likely to have well established identities. Individuals who self-select into these communities are more likely to stay because of superior matching. Hence, we would expect that older communities would retain members at higher rates.

Hypothesis 4 (H4): *Community Age is positively associated with member retention in an online community*.

As communities evolve and newer communities are formed, we would expect the age of a community to have differential effects on member attraction and retention for narrow niche and broad niche communities. Newer members would be more likely to be attracted to narrow niche communities as the discussion would be focused around their areas of interest. However, as communities get older, communities with narrow niches create or evolve very specific norms and processes and newer members might feel more intimidated communicating with existing members. On the other hand, the homophily effects will not be as great in communities with broader niches as the areas of topics covered would be broader and hence we would expect that with increasing age of communities, newer members would be attracted more to communities with broader niches than to communities with narrower niches.

Hypothesis 5 (H5): *Community age positively moderates the effect of member overlap density on member attraction in an online community.*

As communities evolve, we would expect that members in communities of narrower niches would be more homophilous than members in communities with broader niches. This is because with a narrow focus of interest, members of communities with well-defined narrow topics would bond better than with members in communities with topics covering a wide variety of topics. These members would be less willing to leave the community. Hence we would expect that with increasing age, narrow niche communities would retain members better compared to broad niche communities.

Hypothesis 6 (H6): *Community Age negatively moderates the effect of member overlap density on member retention in an online community.*

2.2 Multiple Category Membership and Member Motivation for Contribution and Participation

Member overlap density, a construct at the community level, is equivalent to the average of multiple memberships of its members. For example, if there are three members in community A, where each member is a part of 10, 12, and 14 other communities, and if another community B contains 4 members who are part of 2, 3, 3, and 4 communities then the member overlap density of A is 12 while that of B is 3. I conduct further analysis at the individual level since it is individual member behavior that drives community growth and members who are in fewer communities might behave differently compared to members who are in multiple communities. In this section, I develop theory and hypotheses to understand how individuals with multiple memberships participate and engage on the platform and in individual communities.

2.2.1 Multiple Category Membership

Researchers in economic sociology have studied the association between multi-category participation and its impact on evaluation at the firm, product and individual level (Hsu et al. 2009, Koçak et al. 2009). They argue that a category-membership specialist has a higher actual expected appeal in his or her focal category than any membership generalist (Hannan 2010). Most findings at the firm, product and individual level show that category spanners have lower appeal and are discounted by the audience. This lower appeal has been found in the context of firms (Zuckerman 1999), film actors (Zuckerman et al. 2003), feature films and e-bay auctions (Hsu et al. 2009) and restaurants (Kovács and Hannan 2010).

While prior research on category spanning has focused on the evaluation of actors who span categories, another line of literature on cultural omnivorousness analyzes choices faced by a heterogeneous audience (Goldberg et al. 2016). The engagement of actors (defined as the amount of effort or focus in each category) has not been a particular emphasis of study in prior research. Engagement as an audience member involves learning about producers, evaluating producers and products, constructing labels and schemas, and consuming or discussing products and services (Hsu et al. 2009). Cattani et al. (2008) study consensus about labels from a networks perspective and find that a less engaged and more fragmented audience (less

dense ties, fewer repeat ties with producers and higher turnover) agrees less about category boundaries. Engagement has generally been examined from an audience perspective, with findings indicating that highly engaged audience members develop more subtle and finer grained distinctions; are more likely to develop more elaborate categorical schema and that audiences that spans categories and values variety and typicality are resistant to products that span boundaries (Goldberg et al. 2016). Because boundary spanners play a crucial role in protecting categorical boundaries, it is important for us to understand how these audiences who are variety seeking participate in online communities. In this paper, I study how multi category participation affects the long term participation of members on the platform and on the engagement of members in each category.

2.2.2 Member Motivation for Contribution and Participation

Research on online communities has primarily focused on the motivation of individual members for participating in online communities. Scholars in this area have surveyed participants in online groups to understand their motivation for participating in these groups despite the participants not receiving any monetary compensation. Economic theories suggest that users participate when the tangible and intangible benefits that they accrue are greater than the costs (effort) of participation (Olsen 1965). Extrinsic motivations, intrinsic motivations, tangible benefits and intangible benefits accrued by individuals have been varyingly stated as reasons for user participation in online communities. User benefits such as learning and enhanced reputation are some of the extrinsic motivations and tangible benefits that users gain when they participate in Open Source Software (OSS) projects (Lerner and Tirole 2002). Wasko and Faraj (2000) surveyed 3 technology-focused Usenet communities and found that the primary motivations for user participation included attaining ‘tangible’ returns such as gaining access to valuable information and obtaining answers to specific questions, while also receiving intangible returns such as enjoyment and interaction with the community.

A participant's online and offline reputation is also influenced by his or her experience in the community. Users gain status points for contributing answers on communities such as Stack Overflow. Moreover, extrinsic factors, including career concerns, can explain some users' motivations to participate in online communities (Xu et al. 2014). A user's contributions to Open Source Software and to online technical communities such as Stack Overflow are used by recruiters to assess their technical capabilities thus affecting their career prospects.

User interactions with each other, reciprocity within the community, and a sense of commitment to the community can also explain why users continue to participate in the community. Drawing from theories of organizational commitment, Bateman et al. (2011) distinguish different types of commitment (continuance, affective and normative) and how they affect user behavior (reading threads, posting replies and moderating discussions) in online communities. The initial interaction that the user has with the community is important as it affects the likelihood of user participation in the future (Joyce and Kraut 2006). New members who received a reply to their initial post were more likely to post to the community again. Preece and Ghozati (2001) study empathy in different online communities and find that an overwhelming percentage of communities (81%) had some form of empathetic communication and that empathy was stronger in support communities. People who share common interests and who are similar to each other tend to be more empathic towards each other (Colvin et al. 1997), and such homophily might explain why users continue to participate in online communities.

Most of the online groups do not find regular participants to contribute to the group. Cummings et al. (2002) observe that one-third of listservs do not have any communication over a 130 day observation period and that in the other communities, traffic was low with skewed participation where a few subscribers generated a majority of the messages. Jones et al. (2004) examine 600 Usenet newsgroups over a 6 month period and find that the churn of users in these groups is high as only 11.5% of the people returned to post the next month.

A community's norms, structures, processes, and beliefs evolve as it ages and as members join and leave the community. Communication volume and membership size of a community have been found to have both negative and positive effects on community sustainability (Butler 2001). As communication activity in a community increases, users change their pattern of responses by replying to simpler messages, by writing simpler messages or even by ending participation (Jones et al. 2004).

Most public online communities do not have a selection process for a member to join the community. Some communities, including the Apache server project and Linux, have a meritocracy driven hierarchical system where existing members vote to grant developers more privileges based on their contributions. In Wikipedia, prolific contributors become administrators where they gain extra privileges including the authority to edit protected pages or the ability to block specific users from editing; on Reddit, regular contributors are invited to become moderators of the community. An individual member decides to join and contribute to those forums in which he is interested. Competition among communities and the location of a community on the platform determines how communities differ in their growth. Some communities manage to retain more existing members while other communities have a larger churn, where they do not have a high retention of existing members but gain a large number of new members.

The problem of participation in and contribution to online communities can be broken down into two components: 1) motivation to join the online communities and 2) motivation to continue participation after joining (Joyce and Kraut 2006). At the community level, prior researchers have suggested three models/routes which increase an individual's commitment to a group; a reinforcement model: people repeat actions that lead to positive reinforcements; a reciprocity model: a favor is likely to be returned; and a personal bond model: repeat interactions increase attachment and personal bonds (Joyce and Kraut 2006, Ren et al. 2012). Members who join few communities are more attached to their community and users who are more attached to the community more likely return to the community. An increase in both bond-based and identity-based community level attachment leads to greater participation (Ren et al. 2012).

Considering communities on a single platform, members who participate in multiple communities are interested in multiple topics and are more willing to stay longer on the platform as they can get all their information needs met from that single platform. These members are less likely to be attached to any particular community, but are more likely to be interested in diverse topics and varied subjects. Members subscribe to news feeds from communities of interest, which increases their expected match value: the expected utility of examining the opportunity and interacting with it, while minimizing navigation cost as they can get different feeds from various communities on a single page. Participant overlap in different parts of the platform creates synergies where a personal connection made by two people on the platform also increases the value they get from interacting in another space (Kraut and Resnick 2012). A single platform, containing multiple spaces, allows members to create personal bonds with the same alter in multiple communities, increasing bond based identity. These members are likely to have overlapping interests in similar communities due to the homophilous nature of member interests (Olson and Neal 2015). Members in multiple communities have a higher switching cost since they might have to join multiple platforms to meet their need for varied topics, which would result in creating multiple user ids or profiles and increasing participation costs. Members moving to a new platform have to learn to find areas of communities that are of interest to them (Kraut and Resnick 2012), and these costs are higher for members with varied interests. For the above reasons we expect to find support for the following hypothesis.

Hypothesis 7 (H7): *Multiple Membership is negatively associated with a member's rate of exit from the platform.*

We would expect both multiple membership as well as platform and community engagement to positively affect the duration of stay of a member on the platform and in a community. However, we do not know which of these would have a greater effect and hence I treat this as an empirical question.

Members can choose to join one or more communities among the multiple communities that are present on a platform. The members then participate in these communities by replying to posts or

commenting on articles. The core assumption by prior researchers for engagement of an actor or a producer is generally that each member of a population of producers has the same finite level of resources for engagement (Hsu et al. 2009). Since members are constricted by time and expend effort for their contributions, participation in multiple categories involves a tradeoff. When members belong to multiple communities, then their time and effort is dispersed among multiple communities and hence we would expect these users to have a lower participation in each community. Some members only choose to join and participate in a few communities of interest. Since these members belong to very few communities, their time and effort is not divided among multiple communities. Participation is also affected by the strength of member's identification with the community and the kind of interpersonal bonds that they develop with other members in the community (Ren et al. 2012). Members belonging to fewer communities identify themselves more with the communities that they are a part of and have deeper interactions with other members of those communities. As members increase their multiple community membership, they are likely to increase their overall engagement on the platform but their engagement per community is likely to decrease.

Hypothesis 8 (H8): *Multiple Membership is positively associated with overall engagement on the platform.*

Hypothesis 9 (H9): *Multiple Membership is negatively associated with engagement per community in online communities.*

3. Data

The data are from Reddit.com, one of the largest online communities founded in December 2005. During its early years, Reddit only had two English language communities, which were very flexible on the topics that they allowed users to post. Towards the middle of 2007, Reddit allowed users to create their own communities on any topic, and Reddit saw an explosion in the number of communities from 2008 onward. There were 18,664 authors who contributed to one or more of the communities on Reddit in Jan 2008 while there were 243,797 authors in Dec 2010 (Fig. 1). Similarly, the number of successful communities

(communities that attract 10 or more authors in a month) increased from 26 in Jan 2008 to 1,639 in Dec 2010 (Fig. 1).

Insert Figure 1 about here

The data cover all communities in Reddit from 2006 to 2010 during its early explosive growth phase. It encompasses the number of users Reddit attracted and the number of successful communities that were created on the platform. Any individual can join a public community on Reddit since there is no formal requirement for joining a community. Communities have many lurkers, but they are invisible and so we can only recognize members by their user ids if they post in any of the communities.

The data consists of all the new communities that have been created, comments that have been posted, time of the comment, and authors who have written the comment. The panel data cover the 60 month period from Jan 2006 to December 2010. To test the first set of hypotheses, I consider data from Jan 2008 to June 2010 as I need sufficient number of communities to construct the time varying measure of member overlap density. The last 6 months of the data are used for constructing the measures of member retention and survival.

To test the causal effect of multiple membership on engagement, I consider data from Jan 2006 to December 2008. As there were very few members during the early period, I considered the set of individuals who joined the platform before October 2006 and also survived on the platform till June 2008. Individuals were constrained to only 2 communities and a few minor language specific communities till August 2007 but could quickly increase their multiple membership as newer communities were added to the platform after 2007. This event serves as an exogenous shock as there was a variation in the increase in multiple membership for different individuals after August 2007.

An individual is denoted by subscript ‘i’, a community by subscript ‘j’ and time by subscript ‘t’.

3.1 Dependent Variables

I measure the effect of the outcome at the community and individual levels.

3.1.1 Community level outcomes

Count of New Members$_{jt}$: The number of new members of a community in a month is the number of authors who have posted for the first time in the community during that month. I have considered the author to have become a member of the community at the time of the first post. These are members who have not posted in that community in any of the prior months from the time of creation of the community or from January 2008.

Count of Retained Members$_{jt}$: The number of retained members for a month is the number of members who still remain in the community at the end of each month. These are members who come back to post in the future months. I assume a member to have left the community during the month of the last post if the last post of a member in a community is before June 2010. If a member has not posted in a community in the last 6 months or more then I consider that the member has left the community at the time of the last post. I consider 6 months as an appropriate time frame because very few members return to post in a community after 6 months (Figure 2). An author is considered as a member of a community between the months when he or she made his or her first post and his or her last post in that community.

Insert Figure 2 about here

3.1.2 Individual level outcomes

Overall Engagement$_{it}$: Overall Engagement is defined as the number of comments made on the platform by a member in that month.

Engagement per Community$_{it}$: Engagement per Community is defined as the number of comments made in a community divided by the number of communities that the member belongs to during that month.

Exit from Platform$_{it}$: This is a dummy variable that takes the value of 0 if the member returns to any of the communities on the platform later and takes the value of 1 if he or she quits the platform. For example, if a member posts on any of the communities on the platform for the months of May, June, August, September 2008 but does not post in any of the communities after September 2008 then the value for this measure will be 0 for the months of May, June, July and August 2008 and will be 1 for September 2008. A member is assumed to have joined the platform in the month when the user makes their first comment in any of the communities. If the user has not made any comment during the last 6 months on the platform then the user is assumed to have exited the platform.

Exit from Community$_{it}$: This is a dummy variable that takes the value of 0 if the member is retained in that community for the following month and takes the value of 1 if he or she quits the community. It is constructed similar to the measure of *Exit from Plaform* except that I study if a member is retained in the focal community of interest. A member is assumed to have joined the community in the month when he or she makes his or her first comment in the community. If a member makes a comment in a community before June 2010 and has not made any other comment in that community till Dec 2010 then I assume that the user has exited the community.

3.2 Focal Independent Variables

Member Overlap Density$_{jt}$: I construct member overlap density similar to prior studies in the ecological literature and in online communities (Baum and Singh 1994, Wang et al. 2013). Two communities share a member if the member is a part of both the communities in that month. I have the entire population of public groups on Reddit, and for every community I identified the number of groups on Reddit with which it shares members and also the number of members shared with that group. The degree of membership overlap between a focal community and another community in a particular month is the number of shared members in the two communities in that month. Member overlap density for a community is then calculated as the sum of the degree of membership overlap with all other communities with which the group shares members,

which is then weighted by the total number of members in the focal community in that month. Member overlap density for a community i at time t is given by the following formula, where j are the communities with which the focal community i shares its members.

$$Member\ Overlap\ Density_{it} = \sum_{j=1}^{J} \frac{\#Sharedmembers_{ijt}}{\#Members_{it}}$$

Multiple Membership$_{it}$: Multiple membership for an individual is the number of distinct communities that the individual belongs to during that month.

The description of the dependent variables, focal independent variables and other control variables are shown in Table 1.

Insert Table 1 about here

4 Methods

Descriptive statistics of the data used for the first two regressions are provided in Table 2. All observations are at the community level. Community age was calculated in months and ranged from 2 months to 30 months. Since I used fixed effects regressions at the community level, I considered communities with at least 2 months of data. I also used the last 6 months of data to calculate the number of members retained in the community. Hence any community that was created during or after the 30th month, that is, after June 2010, has not been included in the analysis. There were also many communities with one or very few users that either had no activity or very minimal activity over long periods of time. To ensure that there was some minimum level of activity in the communities, I only considered communities that had more than 10 members. Reddit has some large default communities which continue to grow larger because all users who join Reddit are automatically subscribed to these communities. Hence,

I removed the early and default communities as these may skew the results. As robustness tests, I also ran regressions on the entire dataset, including all early and default communities, and the main results are not affected. The variables measuring the number of members in focal community and the number of members in other communities were log transformed as they were highly skewed.

Insert Tables 2 & 3 about here

To test the proposed hypotheses, I ran multiple specifications. The first two specifications are at the community level to test the effect of member overlap density on the count of new members and the count of retained members respectively. Specifications 3, 4 and 5, at the individual level, test the effects of multiple membership on the exit rate from the platform and community and on the level of engagement.

The first specification predicts the count of new members as a function of member overlap density, while the second specification predicts the count of retained members as a function of member overlap density. I use the fixed effects poisson model to account for unobserved community heterogeneity with robust standard errors.

$$\text{Count of New Members}_{jt} = \alpha_j + \beta_{1*}\text{Member Overlap Density}_{jt} + \gamma_k*\text{Controls} + \varepsilon_{jt} \quad \text{- (1)}$$

$$\text{Count of Retained Members}_{jt} = \alpha_j + \beta_{1*}\text{Member Overlap Density}_{jt} + \gamma_k*\text{Controls} + \varepsilon_{jt} \quad \text{- (2)}$$

where controls are community age, average score of focal community, average level of participation of the focal community, concentration of participation, number of members in the focal community, and number of members in other communities.

The third and fourth specifications are piecewise constant proportional hazards models at the individual level that test the effect of multiple membership on rate of exit from a platform and on rate of exit from a community respectively. The descriptive statistics and correlations for variables used in the third specification are provided in Appendix tables A1 and A2 while those used for the fourth specification are provided in Appendix tables A3 and A4.

$$\text{Exit from Platform}_{it} = \alpha_j + \beta_{1*}\text{Multiple Membership}_{it} + \gamma_k*\text{Controls} + \varepsilon_{jt} \quad \text{- (3)}$$

where the controls are number of comments of the member on the platform, average score of member on the platform and total number of members on the platform.

$$\text{Exit from Community}_{it} = \alpha_j + \beta_{1*}\text{Multiple Membership}_{it} + \gamma_k*\text{Controls} + \varepsilon_{jt} \quad \text{- (4)}$$

where the controls are number of comments of the member in the focal community, number of comments of the member in all the other communities, average score of member in the focal community, concentration of participation, member overlap density, number of members in the community and number of members in all other communities.

The fifth specification is a poisson regression that tests the effect of multiple membership on engagement. The descriptive statistics and correlations table for the fifth model are provided in Appendix tables A5 and A6.

$$\text{Overall Engagement}_{it} = \alpha_j + \beta_{1*}\text{MultipleMembership}_{it} + \gamma_k*\text{Controls} + \varepsilon_{jt} \quad \text{- (5)}$$

$$\text{Engagement per Community}_{it} = \alpha_j + \beta_{1*}\text{MultipleMembership}_{it} + \gamma_k*\text{Controls} + \varepsilon_{jt} \quad \text{- (6)}$$

where the controls are average score of the member, the cumulative number of comments that the member has made on the platform till that month and the number of members on the platform in that month.

5 Results

Hypotheses 1 posits that the effects of member overlap density on the count of new members is negative since communities that have a higher overlap density would have a harder time attracting new members. As shown in model 2, I find that my hypothesis is supported as the coefficient of member overlap density on count of new members is negative and significant ($\beta = -0.072$, $p < 0.01$). Figure 3 displays the rate multiplier for new member attraction for the unit change in member overlap density over the interval from 1 to 51. As member overlap density increases from its minimum value of 1 to its mean of 25.34, the attraction rate falls by 83%. We also observe that community age has a significant negative effect on new member attraction ($\beta = -0.0492$, $p < 0.01$).

Insert Figure 3 and Table 4 about here

Hypothesis 2 posits that the effect of member overlap density on member retention is negative. Surprisingly, as seen in model 5, not only do we find that the hypothesis is not supported, there is a strong significant support in the opposite direction. Contrary to the prediction of hypothesis 2, member overlap density has a positive effect on member retention ($\beta = 0.029$, $p < 0.01$). As member overlap density increases from its minimum value of 1 to its mean of 25.34, the retention rate of a community doubles. Figure 4 displays the rate multiplier for member retention for unit change in member overlap density over the interval 1 to 51. We observe that community age has a positive effect on member retention ($\beta = 0.0194$, $p < 0.01$). Although, older communities attract fewer new members, their retention rate of existing members is higher compared to newer communities.

Insert Figure 4 about here

As seen in model 2, we find the effect of community age on member attraction is negative ($\beta = -0.0492$, $p < 0.01$) supporting hypothesis 3. Older communities find it harder to attract new members compared to newer communities. However, we find in model 5 that community age has a positive effect on member retention (($\beta = 0.0194$, $p < 0.01$) supporting hypothesis 4. Although, compared to newer communities, older communities attract new members at lower rates; they retain their existing members at higher rates.

In models 3 and 6, I test the interaction effects of member overlap density and community age on member attraction and member retention. We find that the interaction effect of member overlap density and community age on member attraction is positive ($\beta = 0.00122$**,** $p < 0.01$) supporting hypotheses 5. The interaction effect of member overlap density and community age on member retention is negative ($\beta = -0.000548$**,** $p < 0.01$), supporting hypothesis 6. However, although statistically significant, the magnitudes of the interaction effects on member attraction ($\beta = 0.00122$**,** $p < 0.01$) and member retention ($\beta = -0.000548$**,** $p < 0.01$) are very small compared to the main effects. We can observe this graphically in Figures 5 and 6. The rate multiplier of new member attraction gradually declines with increase in member overlap density

and community age (Fig 5) while the rate multiplier of member retention gradually increases with increase in member overlap density and community age (Fig 6).

Insert Figures 5 and 6 about here

Figures 7 and 8 depict the survival estimates for members on the platform and in one large community, "AskReddit". There is a high probability (55%) of a member leaving the platform during the first month (Fig. 6). The probability of a member leaving the community "AskReddit" in the first month is around 40% (Fig. 8). The hazard rate of a member leaving the platform or the community levels off after the first month.

Insert Figures 7 and 8 about here

I estimate piecewise constant proportional hazards models on all members to understand the effect of multiple membership on the rate of exit from the platform and from a particular community. For a unit increase in the multiple membership, the hazard rate of a member exiting the platform drops by 26% (Model 9). As robustness tests, I also estimate a fixed effects logistic regression to account for individual heterogeneity and observe that the findings are corroborated. Model 9 shows that is a strong statistically significant negative effect ($\beta = -0.149$, $p < 0.01$) on the exit rate of a member leaving the platform. These results support hypothesis 7.

Similarly, for a unit increase in multiple membership, the hazard rate of a member exiting a community drops by 15% (Model 14). Members who participate in multiple communities are more likely to stay on the platform and are also more likely to stay in their communities compared to members who belong to fewer communities. Comparing figures 9 and 10, I find that the effect of multiple membership on a member exiting a community is high compared to the effect of engagement in the community. As robustness tests, I conduct similar analysis on two other communities and observe the same effect (Appendix Table A7).

Insert Tables 5 and 6 and Figures 9 and 10 about here

I create new variables at the individual level, overall engagement and engagement per community, and test hypotheses 8 and 9 which concern the effects of multiple membership on engagement (Table 7). I estimate a fixed effects poisson regression and find that multiple membership increases the overall engagement of members on the platform as seen in Model 16, supporting the eighth hypothesis ($\beta = 0.024$, $p < 0.01$). I also estimate a fixed effects regression to test the effect of multiple membership on average engagement and find a negative effect, supporting the ninth hypothesis ($\beta = -0.416$, $p < 0.01$).

Insert Table 7 about here

Multiple membership decreases the rate of exit from the platform and also decreases the rate of exit from the community since members of multiple communities stay longer on the platform and hence stay longer in the communities. However, while increasing overall engagement on the platform, multiple membership decreases the average engagement of members in communities as their effort is spread out across multiple communities.

6 Discussion

This paper addresses the paradox as to why communities with high member overlap density grow at a lower rate but survive longer (Wang et al. 2013, Zhu et al. 2014). Prior research on online communities has either been conducted at the group level or at the individual level. I combine analysis at the group and individual level to offer a deeper understanding of long term user behavior and community competition in online communities. I use the ecological framework of competition to examine a community's ability to attract new members and to retain existing members and also conduct analysis at the individual level to understand how members engage in communities and on the platform.

I separate community growth into its two components of member attraction and member retention and test the effect of member overlap density on new member attraction and long term member retention

separately. I use archival data from Reddit and find that high levels of member overlap density are negatively associated with new member attraction. Surprisingly, I also find that high overlap density is positively associated with long term member retention. Member overlap density as the community construct is equivalent to the average of multiple membership of members of that group. Hence, upon further investigating the behavior of members on the platform and in communities at the individual level, I find that members belonging to multiple communities are more likely to stay on the platform compared to members belonging to fewer communities. Further, I exploit an exogenous shock to study the effect of multiple membership on engagement and find that as individuals increase their multiple membership they increase their overall engagement on the platform but their engagement per community decreases.

I make theoretical contributions to the information systems, organizational ecology and multi-category participation literature. At the community level, my findings augment prior studies that have analyzed the effects of member overlap density on community growth and community survival (Wang et al. 2013, Zhu et al. 2014). I find that communities that compete with multiple other communities find it more difficult to attract new members. I add to the multi-category participation literature by studying the effect of multiple membership on engagement and duration of stay on the platform and in a community. At the individual level, I find that multiple membership increases the likelihood of a members stay on the platform and hence in their communities.

Prior researchers have emphasized the community level factors as important metrics for understanding retention in an online community. My results suggest that there are platform level factors such as multiple membership that may have a significant effect on member retention both on the platform and in the communities.

Popielarz and McPherson (1995) study real world online voluntary organizations and there are some interesting contrasting findings compared to this study. They find that voluntary organizations lose fastest those members who are subject to competition from other groups. They also find that there is a high

rate of turnover for those areas of social space where many organizations compete for members. In other words, in real world voluntary organizations, those organizations with broad niche lose members the most. In contrast, I find that in online communities, members who belong to multiple groups and are subject to the maximum competition are more likely to stay on the platform and in their communities as the platform effect dominates. This discrepancy in findings may occur because the effort involved in contributing to an online community is much lower compared to the effort involved in contributing to real world communities.

The findings of this study point to multiple avenues for future research. Faraj and Johnson (2011) study interaction patterns among members in five online communities and find a pattern for direct and indirect reciprocity between members. Future research could study if there are systematic patterns of interactions among members not only within a community, but also across communities. Research on how interaction among members affects their duration of stay in communities and on the platform is also an interesting area to be explored.

Researchers studying innovation in organizations and knowledge transfer across organizations have focused on the importance of boundary spanners in diffusing information (Aldrich and Herker 1977) and the role of weak ties in sharing knowledge (Hansen 1999). In this vein, it would be fruitful to study if members belonging to multiple categories have a unique role in bringing in external information from other areas that might be beneficial to the community.

In the recent past, researchers have argued that the growth of internet has increased political polarization (Farrell 2012; Fiorina and Abrams 2008) and members join communities that align with their views creating “echo chambers” (Adamic and Glance 2005; Conover et al. 2012). It would be interesting to study if broad niche communities have more moderate views and whether members who belong to multiple communities have less polarizing opinions.

The above results have interesting implications for community management by platform owners. For example, a popular online platform, Stackexchange.com, has a beta mode where users can propose and

start communities. Communities are allowed to continue and become full members of the mail platform only if they are able to reach critical mass by attracting a certain number of participants within a limited time frame. If the proposed communities do not meet this requirement, they are disbanded and are not brought onto the main platform. Communities with high member overlap might take a longer time to achieve a certain user size. Even though communities with high member overlap may have less engaged members in the community, these members are more likely to stay on the platform for a longer time, which might be more beneficial for the platform.

The results of the study highlight paradoxical choices that managers might have to make while designing communities. Managers would want to design niche spaces with limited overlap among communities. However, isolated communities with niche topics retain users for shorter periods of time if these users are not interested in other areas of the platform. My results suggest that managers ought to design communities around similar topic areas as members have homophilous interests and users are more likely to participate longer on the platform if they are a part of multiple communities.

Members might find it difficult to navigate through the large number of communities on a platform and this might lead to their losing motivation and not returning to the platform if they do not find communities that match their interests. The user interface should be designed to make it easy for members to search for new communities. Managers would also want to build good recommendation systems to match new members to communities and to recommend new communities of interest to existing members. To increase user retention on the platform, platform owners should design incentives to encourage individuals to sign up for multiple communities.

Fig 1: Growth of Communities and Members on Reddit from Jan 2008 to Dec 2010

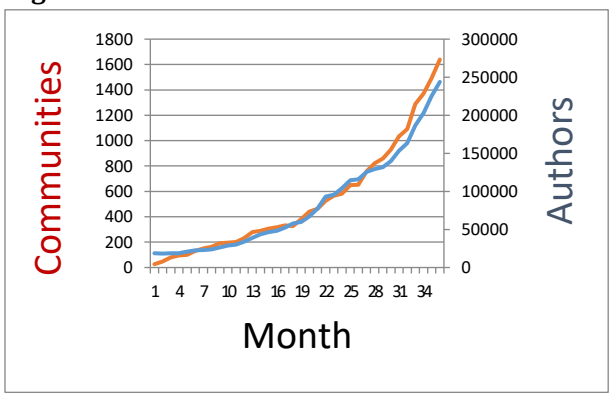


Fig 2: Maximum Time Gap in Months between Two Consecutive Posts by a Member

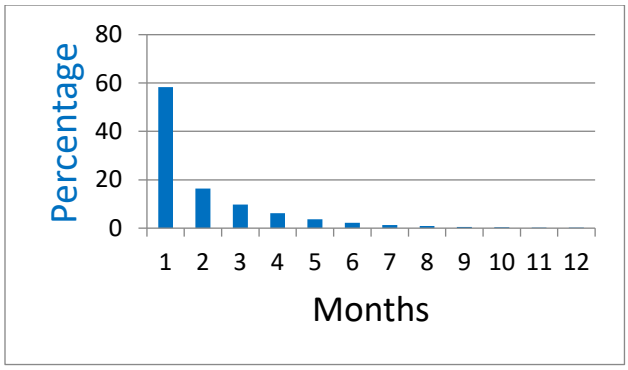


Fig 3: Rate Multiplier for New Member Attraction with Change in Member Overlap Density

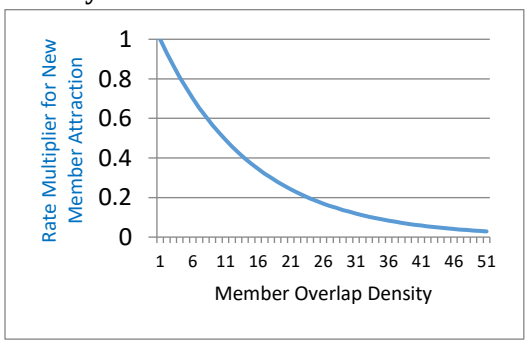


Fig 4: Rate Multiplier for Member Retention with Change in Member Overlap Density

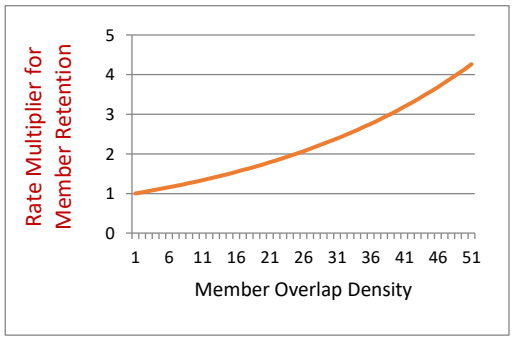


Fig 5: Rate Multiplier of New Member Attraction with Change in Member Overlap Density and Community Age

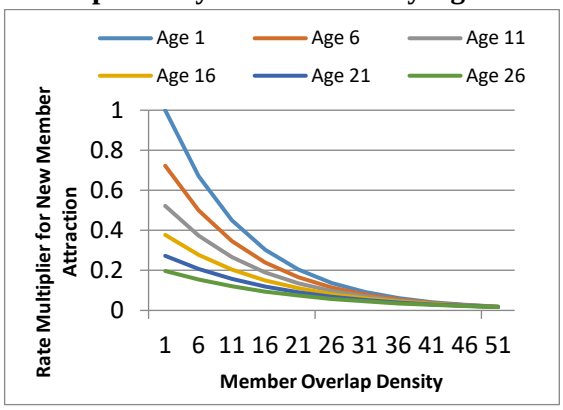


Fig 6: Rate Multiplier of Member Retention with Change in Member Overlap Density and Community Age

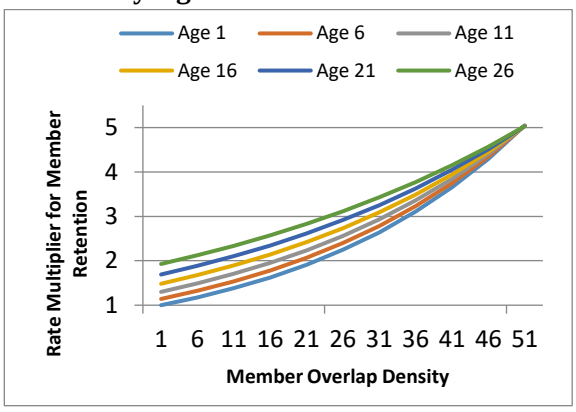


Fig 7: Survival Analysis Graph for Members on the Platform

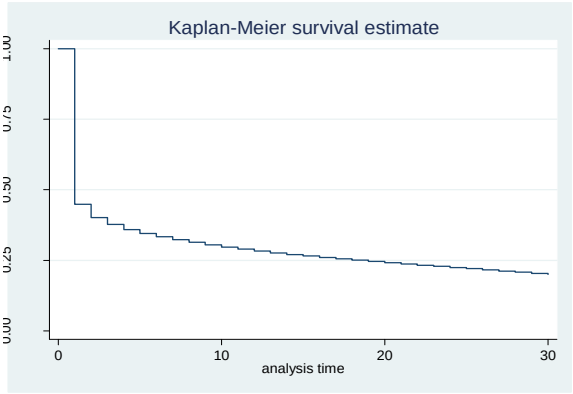


Fig 8: Survival Analysis Graph for Members on Community "AskReddit"

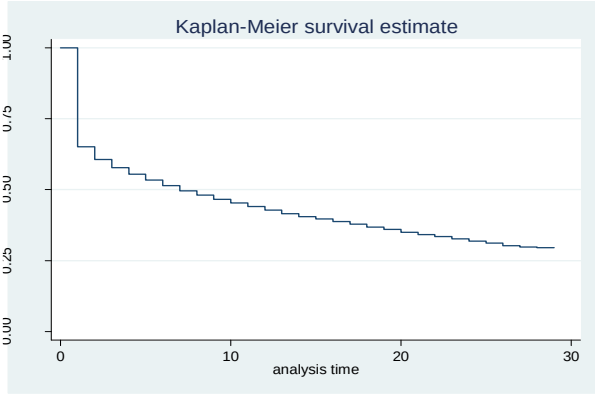


Fig 9: Change in New Member Attraction

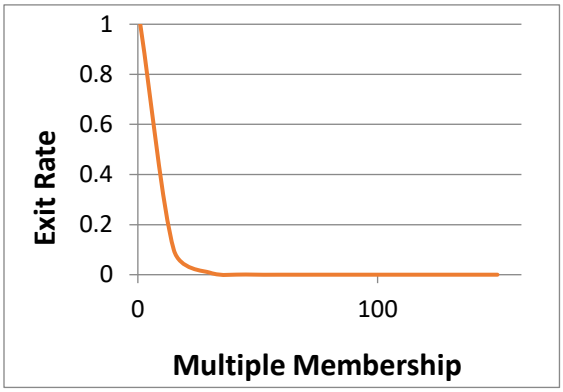


Fig 10: Change in New Member Attraction

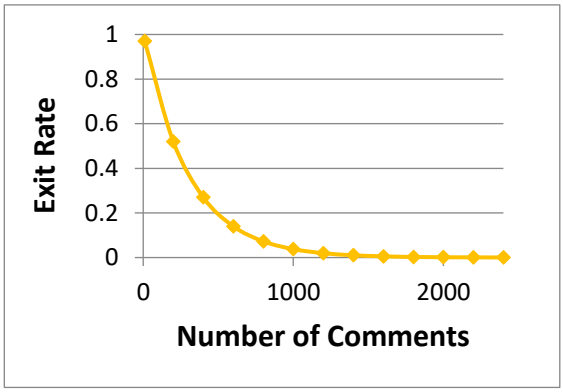


Table 1 Abbreviation of Variables

Variable	Description
# New Members$_{jt}$	Number of new members in a community
# Retained Members$_{jt}$	Number of retained members in a community
Exit from Platform$_{it}$	Dummy variable that takes 1 if a member exits from the platform
Exit from a Community$_{ijt}$	Dummy variable that takes 1 if a member exits from the community
Overall Engagement$_{it}$	Number of Comments made by member on the platform
Engagement per Community$_{it}$	Number of comments made on the platform by a member divided by the number of communities that the member belongs to
Member Overlap Density$_{jt}$	Sum of the degree of membership overlap with all other communities with which the group shares members weighted by the total number of members in the focal community
Community Age$_{jt}$	Age of the community in months
Average Score of Focal Community$_{jt}$	Sum of the scores of all the comments divided by the total number of comments of the focal community
Average Level of Participation$_{jt}$	Total number of comments divided by the total number of members in the community
Concentration of Participation$_{jt}$	Sum of the squares of participation shares of members in a community (HHI index)
# Members in Focal Community$_{jt}$	Number of members in the focal community in that month
# Members in Other Communities$_{jt}$	Difference of number of members on the platform and number of members in the focal community
Multiple Membership$_{it}$	Number of distinct communities that the individual is a member of
# Comments of Member on Platform$_{it}$	Number of comments made by a member on the platform
Avg. Score of Member on Platform$_{it}$	Sum of the score of all comments of a member on the platform divided by the total number of comments
Cumulative Comments of Member$_{it}$	Total number of comments that the individual member has contributed until that month
# Comments of Member in Other Communities$_{ijt}$	Difference in the number of comments that an individual has made on the platform and the focal community
Avg. Score of Member in Focal Community$_{ijt}$	The average score that a member has received for all the comments that he or she has made in the community
# Members on the Platform$_{t}$	Number of members on the platform

Note: *All the variables are constructed at the monthly level*

Table 2 Descriptive Statistics

Variables	Mean	Std. Dev	Min	Max
# New Members$_{jt}$	63.5	84.44	0.00	1256
# Members Retained$_{jt}$	233.39	345.48	0.00	2513
Member Overlap Density$_{j,t-1}$	25.24	7.46	1.00	51.03
Community Age (months) $_{j,t-1}$	13.61	7.21	2.00	30
Avg. Score of Focal Community$_{j,t-1}$	2.13	0.92	-0.33	30.52
Avg. Level of Participation$_{j,t-1}$	2.50	1.75	1.00	34.85
Concentration of Participation$_{j,t-1}$	0.04	0.04	0.002	0.59
# Members in Focal Community$_{j,t-1}$	298.76	415.95	1.00	2862
# Members in Other Communities$_{j,t-1}$	114,006.5	46,895.9	19,936	175,048

N=7,316

Table 3 Correlations

Variables	1	2	3	4	5	6	7	8	9
1 # New Members$_{jt}$	1								
2 # Members Retained$_{jt}$	0.77*	1							
3 Member Overlap Density$_{j,t-1}$	-0.14*	-0.03*	1						
4 Community Age$_{j,t-1}$	0.11*	0.30*	0.16*	1					
5 Avg. Score of Focal Community$_{j,t-1}$	0.38*	0.41*	0.10*	0.06*	1				
6 Avg. Level of Participation$_{j,t-1}$	0.17*	0.14*	-0.21*	-0.008	0.06*	1			
7 Concentration of Participation$_{j,t-1}$	-0.48*	-0.48*	-0.0004	-0.23*	-0.31*	0.07*	1		
8 # Members in Focal Community$_{j,t-1}$	0.76*	0.82*	-0.02*	0.33*	0.43*	0.18*	-0.72*	1	
9 # Members in Oth. Communities$_{j,t-1}$	-0.11*	-0.06*	0.06*	0.42*	-0.06*	0.07*	0.01	-0.10*	1

N = 7,316

* Correlation sig. at $p < 0.05$

Table 4 Effect of Member Overlap Density on New Member Attraction and Member Retention

	Fixed Effects Poisson					
	Number of New Members			Number of Members Retained		
	(1)	(2)	(3)	(4)	(5)	(6)
Community Age$_{j,t-1}$	-0.0185** (0.009)	**-0.0492*** (0.008)**	-0.0662*** (0.008)	0.00545** (0.003)	**0.0194*** (0.002)**	0.0268*** (0.004)
Average Level of Participation$_{j,t-1}$	0.0689*** (0.018)	0.0480*** (0.013)	0.0498*** (0.013)	-0.0251*** (0.006)	-0.0141*** (0.005)	-0.0140*** (0.005)
Average Score of Focal Community$_{j,t-1}$	0.00337 (0.015)	0.0369*** (0.011)	0.0372*** (0.012)	-0.00187 (0.005)	-0.0130*** (0.004)	-0.0130*** (0.004)
Concentration of Participation$_{j,t-1}$	-0.0606* (0.033)	-0.00438 (0.030)	-0.00498 (0.030)	0.0199* (0.010)	0.00798 (0.010)	0.00962 (0.010)
Number of Members in Focal Community$_{j,t-1}$	1.068*** (0.044)	0.968*** (0.034)	1.003*** (0.035)	0.995*** (0.014)	1.051*** (0.013)	1.026*** (0.012)
Number of Members in Other Communities$_{j,t-1}$	-0.326*** (0.124)	0.390*** (0.108)	0.165** (0.075)	0.0586* (0.034)	-0.243*** (0.034)	-0.124*** (0.031)
Member Overlap Density$_{j,t-1}$		**-0.0721*** (0.004)**	-0.0809*** (0.004)		**0.0293*** (0.002)**	0.0329*** (0.002)
Member Overlap Density # Community Age$_{j,t-1}$			**0.00122*** (0.000)**			**-0.000548*** (0.000)**
Observations	7316	7316	7316	7316	7316	7316
No. of Groups	796	796	796	796	796	796
Log Pseudo likelihood	-39459.6	-34855.1	-34518.5	-34679.8	-29881.3	-30577.9

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.0$

Table 5 Effect of Multiple Membership on Member Exiting the Platform

	Piecewise Constant Proportional Hazards Model			Fixed Effects Logistic Regression	
	(7)	(8)	(9)	(10)	(11)
tp1 (month=1)	-0.595***	-4.567***	-4.216***		
	(0.002)	(0.035)	(0.035)		
tp2 (month=2)	-2.268***	-6.150***	-5.706***		
	(0.007)	(0.035)	(0.035)		
tp3 (month=3)	-2.799***	-6.682***	-6.169***		
	(0.009)	(0.036)	(0.036)		
tp4 (month=10)	-3.350***	-7.228***	-6.569***		
	(0.006)	(0.035)	(0.035)		
tp5 (month=15)	-3.798***	-7.667***	-6.844***		
	(0.012)	(0.037)	(0.037)		
tp6 (month=20)	-3.970***	-7.859***	-6.937***		
	(0.016)	(0.039)	(0.039)		
tp7 (month=25)	-4.035***	-7.954***	-6.952***		
	(0.020)	(0.041)	(0.041)		
tp8(month>25)	-3.937***	-7.869***	-6.795***		
	(0.026)	(0.045)	(0.045)		
No. of Comments by Member on the Platform$_t$		-0.0305***	0.00194***	-0.0130***	-0.00424**
		(0.000)	(0.000)	(0.002)	(0.002)
Average score of the Member$_t$		-0.00313***	-0.00162***	0.00784***	0.00814***
		(0.000)	(0.000)	(0.002)	(0.002)
No. of Members on the Platform$_t$		0.354***	0.351***	0.00125***	0.00126***
		(0.003)	(0.003)	(0.000)	(0.000)
Multiple Membership$_t$			**-0.299*** **		**-0.149*** **
			(0.001)		**(0.024)**
Observations	2610071	2610071	2610071	538554	538554
Number of Subjects	539194	539194	539194	82051	82051
Number of Failures	379468	379468	379468		

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6 Effect of Multiple Membership on Member exiting the Community "AskReddit"

	Piecewise Constant Proportional Hazards Model		
	(12)	(13)	(14)
tp1(month =1)	-1.054***	0.549	-1.765***
	(0.004)	(0.607)	(0.611)
tp2(month =2)	-2.674***	-0.925	-3.369***
	(0.012)	(0.617)	(0.622)
tp3(month =3)	-3.049***	-1.035*	-3.486***
	(0.016)	(0.601)	(0.605)
tp4(month =10)	-3.339***	-1.091*	-3.581***
	(0.009)	(0.596)	(0.600)
tp5(month =15)	-3.616***	-1.108*	-3.526***
	(0.023)	(0.624)	(0.629)
tp6(month =20)	-3.730***	-1.373**	-3.740***
	(0.042)	(0.677)	(0.684)
tp7(month =25)	-3.775***	-1.526**	-3.882***
	(0.068)	(0.710)	(0.718)
tp8(month >25)	-3.886***	-1.544**	-3.868***
	(0.209)	(0.753)	(0.761)
# Comments in $Focal_{it}$		**-0.0138*** **	**-0.00328*** **
		(0.001)	**(0.000)**
# Comments in Other $Communities_{it}$		0.000467	0.00669***
		(0.001)	(0.001)
Average Score of Member in $Focal_{it}$		0.00104***	0.00103***
		(0.000)	(0.000)
Concentration of $Participation_t$		-0.0805	0.0431
		(0.093)	(0.092)
Member Overlap $Density_t$		-0.0616***	-0.0566***
		(0.018)	(0.018)
# Members in Focal $Community_t$		-0.0848	0.179**
		(0.084)	(0.085)
# Members in Other $Communities_t$		-0.0922	0.102
		(0.066)	(0.066)
Multiple Membership (# of $Communities)_{it}$			**-0.165*** **
			(0.001)
Observations	814079	814079	814079
Number of Subjects	165440	165440	165440
Number of Failures	84061	84061	84061

Exponentiated coefficients; Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7 Effect of Mutiple Membership on Overall Engagement and Engagement per Community

	Fixed Effects Poisson Overall Engagement		Fixed Effects Regression Engagement per Community	
	(15)	(16)	(17)	(18)
Avg. Score of Member$_{it}$	0.00950***	0.0106***	0.0530***	0.0527***
	(0.002)	(0.002)	(0.007)	(0.007)
Cumulative Comments of Member$_{it}$	0.000104***	0.0000387	-0.00445***	-0.00279**
	(0.000)	(0.000)	(0.001)	(0.001)
Number of members on Platform$_t$	0.463***	0.362***	-0.562**	0.270
	(0.042)	(0.046)	(0.258)	(0.236)
Multiple Membership$_{it}$		**0.0242*** **		**-0.416*** **
		(0.004)		**(0.073)**
Constant			11.48***	4.633**
			(2.268)	(2.091)
Observations	33888	33888	33888	33888
R-square			0.0369	0.0502
No. of Individuals	1548	1548	1548	1548
Log Psuedo likelihood	-252360.1	-248285.7	-116679.9	-116445.7

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

References

Adamic, L.A. and Glance, N., 2005, August. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery* (pp. 36-43). ACM.

Aldrich, H. and Herker, D., 1977. Boundary spanning roles and organization structure. *Academy of management review*, *2*(2), pp.217-230.

Bateman, P.J., Gray, P.H. and Butler, B.S., 2011. Research note—the impact of community commitment on participation in online communities. *Information Systems Research*, *22*(4), pp.841-854.

Baum, J.A. and Singh, J.V., 1994. Organizational niches and the dynamics of organizational mortality. *American Journal of Sociology*, *100*(2), pp.346-380.

Butler, B. S., Bateman, P. J., Gray, P. H., & Diamant, E. I. 2014. An Attraction Selection-Attrition Theory of Online Community Size and Resilience. *MIS Quarterly, 38*(3), 699-728.

Butler, B., Sproull, L., Kiesler, S. and Kraut, R., 2002. Community effort in online groups: Who does the work and why. *Leadership at a distance: Research in technologically supported work*, pp.171-194.

Butler, B.S. and Wang, X., 2012. The cross-purposes of cross-posting: Boundary reshaping behavior in online discussion communities. *Information Systems Research*, *23*(3-part-2), pp.993-1010.

Butler, B.S., 2001. Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information systems research*, *12*(4), pp.346-362.

Cattani, G., Ferriani, S., Negro, G. and Perretti, F., 2008. The structure of consensus: Network ties, legitimation, and exit rates of US feature film producer organizations. *Administrative Science Quarterly*, *53*(1), pp.145-182

Chengalur-Smith, Indushobha; Sidorova, Anna; and Daniel, Sherae L. (2010) "Sustainability of Free/Libre Open Source Projects: A Longitudinal Study," *Journal of the Association for Information Systems*: Vol. 11 : Iss. 11 , Article 5.

Colvin, C. R., Vogt, D., & Ickes, W. 1997. Why do friends understand each other better than strangers do? *Empathic accuracy* (pp. 169-193). New York, NY, US: Guilford Press.

Conover, M.D., Gonçalves, B., Flammini, A. and Menczer, F., 2012. Partisan asymmetries in online political activity. *EPJ Data Science*, *1*(1), p.6.

Cummings, J.N., Butler, B. and Kraut, R., 2002. The quality of online social relationships. *Communications of the ACM*, *45*(7), pp.103-108.

Faraj, S., & Johnson, S. L. 2011. Network Exchange Patterns in Online Communities. *Organization Science, 22*(6), 1464-1480.

Farrell, H., 2012. The consequences of the internet for politics. *Annual review of political science*, *15*, pp.35-52.

Fiorina, M.P. and Abrams, S.J., 2008. Political polarization in the American public. *Annu. Rev. Polit. Sci.*, *11*, pp.563-588.

Goldberg, A., Hannan, M.T. and Kovács, B., 2016. What does it mean to span cultural boundaries? Variety and atypicality in cultural consumption. *American Sociological Review*, *81*(2), pp.215-241.

Hansen, M.T., 1999. The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative science quarterly*, *44*(1), pp.82-111.

Hsu, G., Hannan, M.T. and Koçak, Ö., 2009. Multiple category memberships in markets: An integrative theory and two empirical tests. *American Sociological Review*, *74*(1), pp.150-169.

Johnson, S.L., Safadi, H. and Faraj, S., 2015. The emergence of online community leadership. *Information Systems Research*, *26*(1), pp.165-187.

Jones, Q., Ravid, G. and Rafaeli, S., 2004. Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Information systems research*, *15*(2), pp.194-210.

Joyce, E. and Kraut, R.E., 2006. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, *11*(3), pp.723-747.

Koçak, Ö., Negro, G. & Perretti, F. 2009. *Multiple audiences and categories: actors' careers in films and television in the United States.* Presented at Annu. Meet. Eur. Group Organ. Stud., Amsterdam, June

Kovács, B. and Hannan, M.T., 2010. The consequences of category spanning depend on contrast. In *Categories in markets: Origins and evolution* (pp. 175-201). Emerald Group Publishing Limited.

Kraut, R. E. & Resnick, P. 2012. *Building successful online communities: Evidence-based social design*. Cambridge, MA: MIT Press

Kuilman, J. and Li, J., 2006. The organizers' ecology: An empirical study of foreign banks in Shanghai. *Organization Science*, *17*(3), pp.385-401.

Lerner, J. and Tirole, J., 2002. Some simple economics of open source. *The journal of industrial economics*, *50*(2), pp.197-234.

Macarthur, R. and Levins, R., 1967. The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist*, *101*(921), pp.377-385.

McPherson, M., 1983. An ecology of affiliation. *American Sociological Review*, pp.519-532.

Moon, J.Y. and Sproull, L.S., 2008. The role of feedback in managing the Internet-based volunteer work force. *Information Systems Research*, *19*(4), pp.494-515.

Olsen.M. 1965. *The logic of Collective Action: Public Goods and the Theory of Groups*, Cambridge, MA

Olson, R.S. and Neal, Z.P., 2015. Navigating the massive world of reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science*, *1*, p.e4.

Parks, M.R. and Floyd, K., 1996. Making friends in cyberspace. *Journal of Computer-Mediated Communication*, *1*(4), pp.0-0.

Popielarz, P.A. and McPherson, J.M., 1995. On the edge or in between: Niche position, niche overlap, and the duration of voluntary association memberships. *American Journal of Sociology*, *101*(3), pp.698-720.

Preece, J. & Ghozati, K. 2001. *Observations and Explorations of Empathy Online*. In. R. R. Rice and J. E. Katz, The Internet and Health Communication: Experience and Expectations. Sage Publications Inc.: Thousand Oaks. 237-260.

Ren, Y., Harper, F.M., Drenner, S., Terveen, L., Kiesler, S., Riedl, J. and Kraut, R.E., 2012. Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *Mis Quarterly*, *36*(3).

Sproull, L., M. Arriaga. 2007. *Online communities*. H. Bidogli, ed. Handbook of Computer Networks, Vol. 3. John Wiley & Sons, New York.

Wang, X., Butler, B.S. and Ren, Y., 2013. The impact of membership overlap on growth: An ecological competition view of online groups. *Organization Science*, *24*(2), pp.414-431.

Wasko, M.M. and Faraj, S., 2000. "It is what one does": why people participate and help others in electronic communities of practice. *The Journal of Strategic Information Systems*, *9*(2), pp.155-173.

Wasko, M.M. and Faraj, S., 2005. Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS quarterly*, pp.35-57.

Xu, L., Nian, T. and Cabral, L., 2014. *What Makes Geeks Tick? A Study of Stack Overflow Careers*. Working paper.

Zhu, H., Kraut, R.E. and Kittur, A., 2014, April. The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 281-290). ACM.

Zuckerman, E.W., 1999. The categorical imperative: Securities analysts and the illegitimacy discount. *American journal of sociology*, *104*(5), pp.1398-1438.

Zuckerman, E.W., Kim, T.Y., Ukanwa, K. and Von Rittmann, J., 2003. Robust identities or nonentities? Typecasting in the feature-film labor market. *American Journal of Sociology*, *108*(5), pp.1018-1074.

Appendix

Table A1 Descriptive Statistics

Variable	Mean	Std. Dev.	Min	Max
1 Quit Platform	0.15	0.35	0.00	1.00
2 Multiple Membership (# of Communities)	5.25	7.45	0.00	162.00
3 # Comments by Member on Platform (100's)	0.01	0.05	0.00	3.36
4 Average Score of Member on Platform (100's)	0.00	0.01	-1.14	2.01
5 # Members on Platform (1000s)	110.88	48.55	18.66	175.06

N = 2,610,071

Table A2 Correlations

Variables	1	2	3	4	5
1 Quit Platform$_{it}$	1.00				
2 Multiple Membership (# of Communities)$_{it}$	-0.21*	1.00			
3 # Comments by Member on Platform (100's)$_{it}$	-0.09*	0.54*	1.00		
4 Average Score of Member on Platform (100's)$_{it}$	-0.01*	0.08*	0.04*	1.00	
5 # Members on Platform$_{t}$	0.06*	0.06*	0.02*	0.01*	1.00

N = 2,610,071

Table A3 Descriptive Statistics

Variable	Mean	Std. Dev.	Min	Max
1 Quit Community$_{it}$	0.10	0.30	0.00	1.00
2 Multiple Membership (# of Communities)$_{it}$	9.73	10.03	1.00	148.00
3 # Comments in Focal$_{it}$	26.97	68.06	0.00	2641.00
4 # Comments in Other Communities$_{it}$	21.17	56.78	0.00	2516.00
5 Average Score of Member in Focal$_{it}$	3.33	14.59	-1144.00	1780.00
6 Member Overlap Density$_{t}$	11.91	3.46	7.19	16.74
7 Concentration of Participation$_{t}$	0.01	0.02	0.00	0.04
8 # Members in Focal$_{t}$(1000's)	3.5	8.7	0.081	94.1
9 # Members in Other Communities$_{t}$(1000's)	27.06	9.43	18.5	73.1

N = 814,079

Table A4 Correlations

Variables	1	2	3	4	5	6	7	8	9
1 Quit Community$_{it}$	1.00								
2 Multiple Membership(# of Communities)$_{it}$	-0.22*	1.00							
3 # Comments in Focal$_{it}$	-0.09*	0.52*	1.00						
4 # Comments in Other Communities$_{it}$	-0.09*	0.53*	0.96*	1.00					
5 Average Score of Member in Focal$_{it}$	0.03*	0.02*	0.02*	0.01*	1.00				
6 Member Overlap Density$_{t}$	-0.30*	0.29*	0.09*	0.09*	-0.03*	1.00			
7 Concentration of Participation$_{t}$	0.40*	-0.31*	-0.10*	-0.10*	0.05*	-0.80*	1.00		
8 # Members in Focal$_{t}$	-0.34*	0.42*	0.13*	0.14*	-0.04*	0.80*	-0.89	1.00	
9 # Members in Other Communities$_{t}$	-0.24*	0.43*	0.13*	0.14*	-0.02*	0.69	-0.66	0.88	1.00

N = 814,079

Table A5 Descriptive Statistics

Variable	Mean	Std. Dev.	Min	Max
1 Avg. Comments of Member$_{it}$	5.48	11.33	0	249.83
2 Total Comments of Member$_{it}$	20.37	49.98	0	1648
3 Multiple Membership$_{it}$	3.18	3.17	1	79
4 Average Score of Member$_{it}$	3.58	6	-58	232
5 Cumulative Comments of Member$_{it}$	202.43	478.03	0	12449
6 Number of Members on Platform$_{t}$	13237.45	5070.48	4907	22553

N = 33,888

Table A6 Correlations

Variables	1	2	3	4	5	6
1 Avg. Comments of Member$_{it}$	1					
2 Total Comments of Member$_{it}$	0.80*	1				
3 Multiple Membership$_{it}$	0.08*	0.41*	1			
4 Average Score of Member$_{it}$	0.05*	0.04*	0.07*	1		
5 Cumulative Comments of Member$_{it}$	0.47*	0.75*	0.54*	0.037*	1	
6 Number of Members on Platform$_{t}$	-0.08*	0.10*	0.48*	0.05*	0.28*	1

Table A7 Effect of Multiple Membership on Member exiting communities

	Piecewise Constant Proportional Hazard	
	(1)	(2)
VARIABLES	funny	worldnews
tp1(month=1)	-390.492***	16.618
	(129.109)	(10.997)
tp2(month=2)	-15.837***	-3.249***
	(4.342)	(1.137)
tp3(month=3)	-7.497***	-4.042***
	(1.436)	(1.397)
tp4(month=5)	-7.659***	-4.365***
	(1.458)	(1.464)
tp5(month=10)	-7.598***	-4.375***
	(1.441)	(1.485)
tp6(month=15)	-7.514***	-4.398***
	(1.455)	(1.522)
tp7(month=20)	-7.518***	-4.330***
	(1.481)	(1.542)
tp8(month=25)	-7.327***	-4.030***
	(1.517)	(1.560)
tp9(month>25)	-7.241***	-3.954**
	(1.553)	(1.579)
Multiple Membership (# of Communities)$_{it}$	**-0.110*** **	**-0.063*** **
	(0.001)	(0.001)
# Comments in Focal$_{it}$	0.001***	-0.003***
	(0.000)	(0.000)
# Comments in Other Communities$_{it}$	0.001***	0.001***
	(0.000)	(0.000)
Average Score of Member in Focal$_{it}$	0.002***	0.004***
	(0.000)	(0.000)
Member Overlap Density$_{t}$	701.577***	-69.413
	(232.184)	(43.914)
Concentration of Participation$_{t}$	-0.074***	-0.028
	(0.028)	(0.024)
# Members in Focal$_{t}$	0.610***	0.259
	(0.162)	(0.182)
Observations	632,316	440,336

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Bias in Online Reviews: Variety and Atypicality in Online Gaming

Abstract

In this paper, I study self-selection biases in online reviews which occur because review propensity and review behavior differ across different segments of individuals. Drawing on theories from sociology and marketing and using data from a large online gaming platform which includes ownership, usage and review information, I segment individuals based on their variety seeking and atypicality seeking preferences. I find that poly-mixers - individuals with high variety seeking and atypicality seeking tendencies are less likely to review products but more likely to give lower ratings compared to other individuals. I find that individuals who own more games have higher propensity to review but after controlling for the propensity to review, I do not find any difference in the ratings compared to individuals who own fewer games. I also find that individuals who have more friends on the platform are more likely of giving higher ratings. Further, I find that product usage is a strong indicator of review propensity and individuals are much more likely to review those games that they play the most. Although, majority of the individuals positively review the games that they play the most, surprisingly, I find that a significant proportion of negative reviews (27%) for games are from individuals for whom that game is among their top three most played games.

1. Introduction

In the recent past, there has been an increased interest in studying categories and on how category spanning affects appeal (Zuckerman 1999; Hsu 2006). The general finding is that entities or objects that span multiple categories have lower appeal (atypical objects) compared to objects that span fewer categories (typical objects) and this effect has been found in different industries such as restaurants, movies and books (Hsu 2006; Negro et al. 2010). Many studies have used ratings from online reviews as the focal dependent variable and a proxy measure for product appeal (Hannan et al. 2007; Kovács and Hannan 2010; Kovács and Johnson 2014). Simultaneously, due to the proliferation of social media and user generated content, there has been a growing interest in studying online reviews (Forman et al. 2008; Moe and Shweidel 2012). Although the main focus has been on estimating the impact of reviews on sales (Chevalier and Mayzlin 2005; Duan et al. 2008; Liu 2006) and on the helpfulness of reviews (Mumdabi and Schuff 2010; Yin et al. 2014); studying biases in online reviews has also been an important area of focus (Li and Hitt 2008; Hu et al. 2009). Multiple findings suggest that online ratings are skewed or biased because only the users who are extremely satisfied or dissatisfied with the product opt into review while the majority of the users with moderate opinion are far less likely to review and rate products (Hu et al. 2009; Hu and Li 2011). Prior studies on category spanning do not take into account these biases while estimating the impact on product appeal. In this study, I control for one type of bias – review propensity and show why controlling for review propensity is important as estimates from models that control for review propensity differ from those that don't control for review propensity.

Multiple studies have documented J shaped distribution for online reviews (Li and Hitt 2008; Hu et al. 2009; Hu and Li 2011). On a scale of 1 to 5 (1 being the lowest and 5 being the highest), majority of the users give product ratings of 4 and 5 and some users give ratings of 1. The proportion of users who rate products as 2 or 3 is very low. This is in contrast to observations from experimental evidence where if all subjects are mandated to rate products, the distribution of the ratings are unimodal or normal (Hu et al. 2005). The J shaped distribution of online reviews has been attributed to self-selection biases (Li and Hitt

2008; Hu et al. 2009; Hu and Li 2011). Although, in a majority of online sites, any individual can provide a review, only few consumers opt in to provide reviews causing self-selection biases. Consumers with extreme preferences are more likely to review a product with more neutral consumers opting out from reviewing a product altogether (Koh et al. 2010).

There may be biases in online reviews because different segments of consumers differ in their review propensity and also differ in the average ratings that they provide. Second, heterogeneity in product type is a major focus in economic sociology and suggest that product characteristics such as their typicality would affect their appeal (Goldberg et al. 2016). However, these studies have not accounted for difference in review propensity across product types. This is mostly because although online reviews are easily available and can be scraped from websites, consumer purchase data is not as readily available.

Goldberg et al. (2016) suggest that the appeal for typical or atypical objects can vary across consumer segments based on individual proclivity for variety seeking and atypicality seeking. Although product type, such as the difference in reviews between blockbuster and niche products, has been considered in online reviews (Dellarocas et al. 2010), there have been no studies that have considered both consumer heterogeneity and product type together. In this paper, I consider review propensity for different product and consumer types as they might provide deeper insights to any biases that may occur.

Drawing on prior work (Goldberg et al. 2016), I categorize consumers into four segments based on their variety seeking and atypicality seeking tendencies. I find that poly-mixers – individuals who seek high variety and high atypicality are, on average, less likely to review but more likely to give lower ratings compared to other segments. I also find that the social network of a user is an important predictor of an individuals' review propensity. Individuals who have more friends on the platform are more likely to review compared to individuals who have fewer friends and these individuals are also more likely to give higher ratings. Further, I find that individuals who own more products are more likely to review but after controlling for their propensity to review, I find that there is no difference in the average ratings for

individuals who own more products compared to individuals who own fewer products. Finally, I also find that product usage is a very important predictor of an individuals' review propensity. Individuals who use the product the most are much more likely to review the product compared to individuals who purchase but do not use the product as much.

The rest of the paper is organized as follows: In the next two section, I review the literature on categories, omnivorousness and online reviews and state my hypotheses. In Sections 4, 5 and 6, I describe the data, explain the methods for calculating the focal constructs and describe the variables of interest respectively. In Section 7 and 8, I explain the estimation methods and the results. Section 9 contains the conclusion and practical implications.

2. Literature Review

2.1 Categories and Category Spanning

Studies on categories focusing on the effect of category spanning on appeal have shown a wide range of contradictory findings (Kovács and Johnson 2013, Kovács and Hannan 2015). However, the general finding is that objects that span categories have lower appeal (Zuckerman 1999; Hsu 2006; Negro; Hannan and Rao 2010) and the negative consequences of crossing boundaries is more severe when the categories spanned are distant (Kovács and Hannan 2015). Both producer and consumer side effects have been suggested as reasons for why category spanning affects appeal. The "generalist vs. specialist" argument is that firms or individuals have a finite amount of resources or attention, and thus those who span multiple categories have to divide their resources and attention among all the categories and cannot focus on any one of them (Hsu 2006). Hence the average quality of the offerings of these category spanning firms or individuals is lower compared to those of the specialized ones. Another reason could be that audiences believe generalists have a lower skill compared to specialists. Therefore, audiences punish actors who span categories and assign them lower value. Finally, audiences might also be confused by objects that span

multiple categories. They might find objects belonging to single category more appealing than objects spanning multiple categories (Hannan et al. 2007).

There have been other studies that document that spanning categories is advantageous. Paolella and Durand (2005) theorize that category spanning effects are contingent on clients' theory of value. They argue that for complex problems, non-recurrent issues and/or issues involving high financial stakes, clients prefer producers who are category spanners. They study corporate legal services in three markets and find that category spanners receive better evaluation. Leung and Ng (2014) argue that spanning behavior can be advantageous or disadvantageous depending on other indicators and find that spanning more categories is advantageous for highly reputable applicants. Kovács and Johnson (2013) argue that high quality organizations can benefit from being atypical. Markides and Williamson (1994) find that spanning related categories is beneficial for firms. Alvarez et al. (2005) argue that optimal distinctiveness is relevant for creative industries and internationally renowned film directors span multiple genres to guard their characteristic styles from isomorphic pressures in their field.

The number and type of categories spanned can affect how audiences perceive the object (Hannan et al. 2007; Kovács and Hannan 2010). Kovács and Hannan (2010) find that organizations that span more focused categories (high contrast categories) are devalued more than organizations that span less focused categories (low contrast categories). Much of the research on categories has focused on the heterogeneity in product type while considering the audience evaluating the product as homogenous (Zuckerman 1999; Hsu 2006; Hannan et al. 2007; Kovács and Hannan 2010; Negro, Hannan and Rao 2010). A recent study (Goldberg et al. 2016) considers heterogeneity in audience types and their evaluation of different product types.

2.2 Variety Seeking, Omnivorousness and Atypicality Seeking

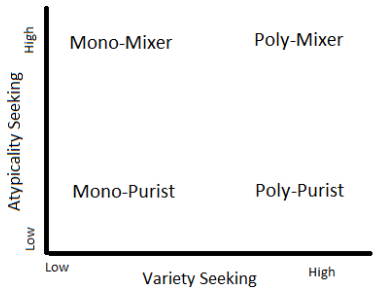
Variety seeking behavior of an individual is an individual's inherent propensity to seek for variety. Some consumers have a high propensity for variety while other consumers exhibit low variety-seeking

behavior (Givon 1984). Explanation for varied behavior of individuals can be classified into two classes: derived motivation - varied behavior is the result of some other motivation; and direct motivation - variation is a motivation of in and of itself (McAllister and Pessemier 1982). Differences in intrapersonal motivations (a type of direct motivation) can largely explain why individuals vary in their proclivity for variety seeking. McAllister and Pessemier (1982) proposed three intrapersonal factors: desire for the unfamiliar, desire for alteration among familiar alternatives (Farquhar and Rao 1976) and desire for information (Raju 1980) to explain differences in variety seeking behavior across individuals. An individuals need for novelty and change are inherently satisfying (Driver and Stuefert 1964; Venkatesan 1973) and individuals may seek out different products to overcome satiation.

Sociologists referring to variety seekers as omnivores, posit that omnivores appreciate a broad variety of genres (Peterson 1992) and omnivorousness is a reflection of cultural openness (Ollivier 2008). Omnivorousness refers to being interested in a wide array of activities or having broad tastes (Peterson 2005) and is generally defined as the number of genres or activities that people practice (Olliver 2008). Omnivores are open minded (Corbin 1980; Novak and Mather 2007; Seale and Rapoport 1997). DiMaggio (1987) proposes that persons with wide ranging networks develop tastes for the widest variety of cultural forms. Relish (1997) suggests that omnivorous taste is also a part of social composition and finds that individuals with a wider social composition are more omnivorous in their tastes. On the other hand, univores display tastes for a narrow range of activities and do not sample widely (Peterson 1992).

Goldberg et al. (2016) further categorize omnivores into three segments based on their proclivity to sample and appreciate atypical objects. Atypicality seekers do not stick to established genres and search for products that are span multiple genres. These individuals actively seek products that combine elements from dissimilar genres in unconventional ways. On the other hand, purists seek out more conventional products. Univore purists are mono-purists as they stick to conventional products from one genre while omnivorous purists are poly-purists as they seek products from multiple genres but do not widely sample

atypical objects. Individuals who stick to a single genre but seek out atypical objects are mono-mixers while individuals who seek atypical products across multiple genres are poly-mixers.



With the recent proliferation of websites containing online reviews, many research studies focusing on category structure or heterogeneity of consumers have used ratings from online reviews (Kovács and Hannan 2015; Kovács and Johnson 2014; Goldberg et al. 2016). In the following section, I detail the literature on online reviews.

2.3 Online Reviews

While the main line of research on online reviews has been on studying the effect of online reviews and ratings on product sales (Chevaliar and Meyzlin 2006; Archak, Ghose and Ipierotis 2011), biases in online reviews is an important area of research (Li and Hitt 2008; Hu et al. 2009). Product ratings reviewed on a scale from 1 to 5 generally show a J shaped distribution (Hu et al. 2009). This is in contrast to controlled experiments where if all users are mandated to provide ratings, online ratings show a unimodal or normal distribution (Hu et al. 2009). This difference between experimental and natural results has been theorized to be driven by self-selection biases. Consumers who highly appreciate the product or consumers who are extremely disappointed by the product are more likely to write reviews while consumers who are neutral about the product are far less likely to write a review (Hu et al. 2009; Koh et al. 2010). Koh et al. (2010) also find that self-selection biases can vary across cultures and find that Chinese consumers are more likely to write reviews compared to US consumers. It has also been observed across different settings that average

ratings decline over time (Duan et al. 2008; Li and Hitt 2008; Zhu and Zhang 2010). This decline in average ratings over time has also been attributed to self-selection bias as early consumers and late consumers have different preferences for products (Li and Hitt 2008). Early consumers are more avid fans of the product and are more likely to give the product positive ratings.

Although product type has not been a major focus of study in online reviews, some researchers have also considered product type while studying online reviews. Dellarocas et al. (2010) find a curvilinear relationship between reviews and blockbuster and niche products. Mudambi and Schuff (2010) find that the effect of review depth on helpfulness is greater for search goods than for experience goods.

Goldberg et al. (2016) consider ratings from reviews but do not account for propensity of consumers to opt into reviewing products. Among consumers who have purchased products, only certain consumers opt into writing online reviews and these consumers review only a few products among the entire set of products that they have purchased. Second, segments of consumers differ across the average ratings that they give for different products. Hence, there is a need to incorporate review propensity while studying consumer heterogeneity, category structure and appeal to mitigate confounding effects from self-selection biases in online reviews.

3. Hypotheses

Mono-purists and mono-mixers consume typical products from their favorite genre. They are more passionate about products that are released in that genre and are more likely to be knowledgeable about that genre and might also want to share their opinion. Poly-purists purchase products across multiple genre but seek only typical products. Poly-purists are the boundary keepers (Goldberg et al. 2016) and might be more willing to state their opinion.

Poly-mixers seek a wide range of products. They are avant-gardes who are open to trying products across multiple genres and which are very atypical. They do not have specific tastes and might come across

products with which they are more likely to be disappointed. Hence poly-mixers might be more likely to review these products negatively.

Hypothesis 1a: Mono-purists, mono-mixers and poly-purists are more likely to review products compared to poly-mixers.

Hypothesis 1b: Reviews of mono-purists, mono-mixers and poly-purists have higher ratings compared to reviews of poly-mixers.

Prior research in public goods shows that peer influence and peer valuation are important factors that affect an individual's propensity to participate and contribute to social goods (Lerner and Tirole 2001; Bagozzi and Dholakia 2006). Muchnik et al. (2013) study social influence biases in online communities and find that opinions of friends or enemies affect ratings given by individuals. Zhang and Zhu (2011) exploit a natural experiment on Wikipedia and find that individuals decrease their contributions when peer group size reduces. Contributors receive social benefits which increase with group size. Individuals with wider social networks are more extroverted and hence more likely to express their opinion (Feiler and Kleinbaum 2015). Individuals might offend one of their friends inadvertently if they review products negatively that their friends appreciate. Hence, we would expect that on average, ratings by individuals who have more friends on the platform would more likely be higher compared to reviews by individuals with fewer friends.

Hypothesis 2a: Individuals with more friends on the platform are more likely to review products.

Hypothesis 2b: Ratings by individuals who have more friends on the platform are higher compared to ratings by individuals who have fewer friends.

Individuals who own more products on a platform are more invested in the platform. These individuals are also more likely to spend more time on the platform compared to individuals who own fewer products. We would expect these individuals to be more actively engaged on forums and chat rooms on the platform and also write more reviews compared to individuals who own fewer products. Elberse (2008)

studies purchase of hit and niche movies and finds that individuals who purchase large number of products venture into the tails to purchase niche products while individuals who purchase few products mostly purchase hit products. Since the average ratings of niche products is lower than that for hit products, I predict that the average ratings given by individuals who own more products should be lower compared to the average ratings given by individuals who own fewer products.

Hypothesis 3a: Individuals who own more products on the platform are more likely to review products.

Hypothesis 3b: Ratings by individuals who own more products are lower compared to ratings by individuals who own fewer products.

Usage of a product is an important indicator of an individual's appreciation of the product. We would expect that individuals are more likely to review products that they use the most and are also more likely to review these products favorably.

Hypothesis 4a: Individuals who use a product more on the platform are more likely to review the product.

Hypothesis 4b: Ratings by individuals who use a product more are higher compared to ratings by individuals who do not use the product as much.

4. Data

The empirical setting for this research is the online gaming industry. The online gaming industry is a relatively new and rapidly growing industry. The data is from Steam which is an online platform owned by Valve Corporation that allows independent developers to host their games. Steam provides a platform for many smaller online gaming publishers and independent developers to reach out to the wider gaming community. Steam hosts over 4,000 games including many popular games such as Dota 2, Team Fortress 2, Doom and Counter-Strike from different publishers and developers.

Steam provides multiple API's which were used to collect data both at the individual level and at the game level. At the game level, I have data on when a game was released, the price of the game, the

reviews for the game, the number of reviews of the game, the percentage of positive reviews for the game and the different tags that are associated with the game. Online video games also have committed fans who tag games with their appropriate attributes. These tags are user-defined and are generally the categories that the game belongs to or the attributes associated with the game.

At the individual level, I have data for around 130,000 individuals on when the individual has joined the platform, the list of all the games that the individual owns and the total number of hours played on each game. I have also collected all the reviews posted by all the individuals across all the games on Steam. Only individuals who have purchased a game are allowed to post reviews on steam. This alleviates the concern of fake reviews that could be a problem in other online review studies where any anonymous individual can post a review. Second, for these users, I have data on all the games that they own as well as the games that these users opt in to review. Very few research studies on online reviews have used both ownership data and review data and thus this study provides a novel setting for studying self-selection bias in online reviews.

5. Calculating Distance Measures

On Steam, games are tagged by users. These tags[1] can be the genre that the users think the game belongs to (Action, Adventure, Role playing) or attributes of a game (multiplayer, individual, co-op, free). For a tag to be visible on Steam, at least 5 different users should have tagged the game using that tag. For any game, a maximum of 20 tags are visible. These are displayed in the descending order of the count of individuals who have attributed the game with that tag. A large game with many gamers playing the game may have many tags while smaller games where only few players play the game might only have 1 tag as there are not as many users tagging these games. Hence, I only consider games with 5 or more tags as games with few tags might not be well defined.

[1] Labels is the more generic notation for tags. I use tags and labels interchangeably. Similarly objects and games are used interchangeably in my study.

Recent studies (Kovács and Johnson 2014, Kovács and Hannan 2015, Goldberg et al. 2016) have analyzed genres to account for structure of genres and incorporate the distance between them in the socio-cultural space. Two categories are related if they frequently co-occur (Gardenfors 2004; Widdows 2005; Kovács and Hannan 2015). Tags that rarely co-occur belong to two different ends in the cultural space (Lizardo 2014; Pachucki and Breiger 2010) while tags that frequently co-occur are seen as being very close to each other. On Steam, many games are tagged as both *Horror* and *Survival* while very few games are tagged as both *Sports* and *Horror*. *Horror* and *Survival* are two tags that are very close to each other while *Sports* and *Horror* lie far apart. Similar to (Kovács and Hannan 2015; Goldberg et al. 2016), I use co-occurrence of tags to map out the relationship among categories and also calculate distance among categories.

Jaccard (1901) proposed a simple category similarity also that takes into account the prevalence of categories. To find the distance between two tags, I initially calculate the Jaccardian similarity measure for the two tags. Let i denote the extension of l_i, the set of objects labeled as l_i, then the Jaccardian similarity measure for tags l_i & l_j is given by $\frac{i \cap j}{i \cup j}$, where i∩j is the number of games in which two tags appear together, while i∪j is the total number of games in which either of the tags l_i or l_j appear where i and j are the set of objects labeled as l_i & l_j respectively.

$$J(i,j) = \frac{i \cap j}{i \cup j}$$

The above Jaccardian index takes a value between 0 and 1. For example, if two tags never appear together then the Jaccardian similarity value between the two tags is 0 denoting perfect dissimilarity. If two tags always appear together then the Jaccardian similarity value between the two tags is 1 which denotes perfect similarity. Following Shepard (1987), Goldberg et al. (2016) posit a negative exponential relationship between perceived sociocultural distance and similarity and the distance between two tags l_i & l_j is given by

$$d(i,j) = -\frac{\ln(J(i,j)}{\gamma}$$ [2]

For example, in my sample, 387 games are tagged as *Horror* while 855 games are tagged as *Simulation*. 61 games are tagged as both *Horror* and *Simulation*. Hence, the Jaccardian similarity between the two tags is 0.051 and the distance between the two games is 5.92.

5.1 Atypicality Measure for a Game

An object is atypical, if it contains many attributes that lie far apart in the cultural space. The further apart these attributes, the more atypical the game to any of the categories. In gaming, *Horror* and *Survival* are two tags that are very close to each other in the cultural space (they co-occur very often) while *Horror* and *Sports* are very far apart. A game tagged as both *Horror* and *Sports* is a very atypical game while a game tagged as *Horror* and *Survival* is a more typical game. Hence, I use the distance between the attributes of objects derived from their frequency of co-occurrence to calculate atypicality measures for objects.

To calculate the atypicality measure for each game, I followed the approach by Goldberg et al. (2016). I only considered games with 5 or more tags. If a game has more than 5 tags, I considered the 5 most popular tags associated with the game and measure the distance between each of the tags to every other tag. Let l(i, x) equal 1 when the tag i is applied to a game x. Let $l_x = \{i|l(I,x) = 1\}$ denote the set of tags applied to the game x. The sum of the pairwise distances between a tag and every other tag of a game gives the overall distance.

$$D(x) = \Sigma_{j\epsilon lx}\Sigma_{j\epsilon lx}\, l(i,x)l(j,x)d(i,j)$$

If the distance measure for a game is very high, it indicates that all the attributes of the game lie very far apart in the conceptual space indicating that the game is very atypical. On the other hand if the distance measure is low, then the attributes of the game are very close to each other and co-occur frequently indicating a typical game. Atypicality of an object ranges from 0 to 1 (Kovács and Hannan 2015; Goldberg et al. 2016). Hence, for this study, I normalize the inverse of the distance measure to ensure atypciality

[2] I have assumed γ to be 0.5. Also, if two tags do not co-occur even in one game then their similarity measure is 0. I have assumed i ∩ j in those cases to be 0.5 as the logarithm of 0 is undefined.

scores range between 0 and 1. 0 indicates that the game is very typical while 1 indicates that the game is very atypical. The histogram of the atypicality index of all the games is given in Figure 1.

5.2 Variety Seeking and Atypicality Seeking Measure for an Individual

I follow the approach by Goldberg et al. (2016) to construct the variety seeking measure for the individual. This method uses a variant of Hausdroff measure and also considers the distance between games in the socio-cultural space to construct the variety seeking measure. An individual who purchases games that lie very close in the socio-cultural space has a low variety seeking propensity compared to a user who purchases games that are very far apart in the socio-cultural space.

To calculate the measure for variety, I initially compute distance between two games by considering the distance between the tags associated with the game. Similar to the construct of atypicality, I consider games with 5 or more tags and only consider the top 5 tags for these games. For any two games A and B, I consider the minimum distance between every tag of game A to game B. Suppose the tags for game A are denoted by the set {a1,a2,a3,a4,a5} and the tags for game B are denoted by the set {b1,b2,b3,b4,b5}, the distance of the closest tag in game B to a1 is denoted this distance is denoted d1. For example, if among the set of tags {b1,b2,b3,b4,b5}, b4 is the closest to a1 then the distance between b4 and a1 is d1. The minimum distance from the tags of B to tag a2 is denoted by d2. Similarly, the minimum distances from B to tags a3, a4 and a5 are denoted by d3, d4 and d5 respectively. The average of d1 to d5 is denoted by h'(A,B).

$$h'(A,B) = \frac{1}{|A|}\sum_{a\in A} \min(d(a,B))$$

Reciprocally, I also estimate the distances from every tag in B to game A and compute the average of the distances and denote this by h'(B,A). Note that h'(A,B) is not necessarily identical to h'(B,A). The average of h'(A,B) and h'(B,A) is the distance between two games and this measure is calculated for every pair of games in the dataset.

$$H'(A,B) = avg(h'(A,B),h'(B,A))$$

The variety-seeking measure of an individual is estimated by computing the average of the pairwise distance between all the games owned by the individual. Individuals who own games that lie far apart in the cultural space have high variety-seeking values, while individuals who own games that are close to each other have low variety-seeking values. The variety seeking measure of an individual y who owns the set of games S is given by

$$\text{VS(y)} = \frac{1}{(n)(n-1)/2} \sum_{x \epsilon S} \sum_{x' \epsilon S} H'(x, x')$$

where n is the number of games owned by the individual.

The variety-seeking measure is invariant to the number of games owned. If the variety seeking index of an individual is very high, then the individual is omnivorous. Analogously, if the variety seeking value is low then the individual is a univore. Figure 2 displays the histogram of the variety seeking index.

The atypicality seeking measure of an individual is the mean of atypicality values of all the games owned by the individual.

6. Variables

$Review_{ij}$: This is a binary variable. This takes the value 1 if an individual i has reviewed game j; else it is 0.

$Recommend_{ij}$: On steam, an individual who reviews a game can give the game either a thumbs up or thumbs down denoting a positive or negative review. If the individual rates a game positively, then the value is 1 but if the individual reviews and rates the game negatively then this value is 0.

Review Pos Neg_{ij}: This variable takes a value of 0 if an individual owns a game and does not review it. It is 1 if an individual reviews the game negatively and is 2 if the individual reviews the game positively.

Playtime of individual for a $game_{ij}$: Overall number of minutes an individual plays a game.

Variety $Seeking_i$: This variable captures the extent of an individual's variety seeking propensity.

Atypicality Seeking$_i$: This variable captures the extent of an individual's atypicality seeking propensity.

Count of Friends$_i$: The number of friends that the individual has on the platform.

Count of Games Owned$_i$: Count of games owned by the individual on the platform.

Days since Join$_i$: Number of days since the individual joined the platform (in 1000's).

Average Playtime of Individual$_i$: Average playtime in minutes for an individual across all the games that the individual owns.

Days since Product Release$_j$: Number of days since the product has been released (in 1000's).

Atypicality of Game$_j$: This variable captures the atypicality value for a game. Atypicality values range from 0 to 1. Scores close to 0 indicate that the game is very typical while games with atypicality scores close to 1 are very atypical.

Average Playtime of Game$_j$: Average number of minutes that the game has been played for across all individuals.

Game Sales$_j$: Number of individuals who own the game in the dataset.

Days since Review Posted$_j$: Number of Days since the review has been posted (in 1000s).

7. Estimation

Model 1 is a binary logistic regression that estimates the effect of individual and products characteristics on positive and negative recommendations without adjusting for review propensity.

$$Recommend_{ij} = \alpha + \beta_1 * Count\ of\ Games\ Owned_i + \beta_2 * Count\ of\ Friends_i + \beta_3 * Days\ since\ Join_i + \beta_4 * Average\ Playtime\ of\ Individual_i + \beta_5 * Overall\ Playtime_{ij} + \beta_6 * Average\ Playtime\ of\ Game_j + \beta_7 * Atypicality\ of\ the\ Game_j + \beta_8 * Game\ Sales_j + \beta_9 * Price\ of\ the\ Game_j + \beta_{10} * Days\ since\ Release_j + \beta_{11} * Days\ since\ Review\ Posted_{ij} + \beta_{12} * Mono\ Purist_i + \beta_{13} * Poly\ Purist_i + \beta_{14} * Mono\ Mixer_i \quad - (1)$$

Model 2 is a Two Stage Heckman Probit regression where the second stage estimates the effect of individual and product characteristics on positive and negative recommendations after the first stage estimates the effect of individual and product characteristics on review propensity.

$$Recommend_{ij} = \alpha + \beta_1 * Count\ of\ Games\ Owned_i + \beta_2 * Count\ of\ Friends_i + \beta_3 * Days\ since\ Join_i + \beta_4 * Average\ Playtime\ of\ Individual_i + \beta_5 * Overall\ Playtime_{ij} + \beta_6 * Average\ Playtime\ of\ Game_j + \beta_7 * Atypicality\ of\ the\ Game_j + \beta_8 * Game\ Sales_j + \beta_9 * Price\ of\ the\ Game_j + \beta_{10} * Days\ since\ Release_j + \beta_{11} * Days\ since\ Review\ Posted_{ij} + \beta_{12} * Mono\ Purist_i + \beta_{13} * Poly\ Purist_i + \beta_{14} * Mono\ Mixer_i$$

$$(Review_{ij} = \alpha + \beta_1 * Count\ of\ Games\ Owned_i + \beta_2 * Count\ of\ Friends_i + \beta_3 * Days\ since\ Join_i + \beta_4 * Average\ Playtime\ of\ Individual_i + \beta_5 * Overall\ Playtime_{ij} + \beta_6 * Average\ Playtime\ of\ Game_j + \beta_7 * Atypicality\ of\ the\ Game_j + \beta_8 * Game\ Sales_j + \beta_9 * Price\ of\ the\ Game_j + \beta_{10} * Days\ since\ Release_j + \beta_{11} * Mono\ Purist_i + \beta_{12} * Poly\ Purist_i + \beta_{13} * Mono\ Mixer_i) \quad -(2)$$

Model 3 is a Multinomial Logistic Regression which estimates the effect of individual and product characteristics separately for positive and negative reviews compared to the base case of not reviewing the product at all.

$$Review\ Pos\ Neg_{ij} = \alpha + \beta_1 * Count\ of\ Games\ Owned_i + \beta_2 * Count\ of\ Friends_i + \beta_3 * Days\ since\ Join_i + \beta_4 * Average\ Playtime\ of\ Individual_i + \beta_5 * Overall\ Playtime_{ij} + \beta_6 * Average\ Playtime\ of\ Game_j + \beta_7 * Atypicality\ of\ the\ Game_j + \beta_8 * Game\ Sales_j + \beta_9 * Price\ of\ the\ Game_j + \beta_{10} * Days\ since\ Release_j + \beta_{11} * Mono\ Purist_i + \beta_{12} * Poly\ Purist_i + \beta_{13} * Mono\ Mixer_i \quad -(3)$$

8. Results

Table 3 displays the results of Model 1 which estimates the effect of product and individual characteristics on recommendation without adjusting for the review propensity. We find that mono purists, poly purists and mono mixers do not differ in their average recommendations compared to poly mixers. We find that individuals who own more games are more likely to provide lower ratings compared to individuals who own fewer games. We find that individuals who have more friends on the platform are more likely to provide positive ratings compared to individuals who have fewer friends. We also find that the overall playtime of the individual for the game has a very significant effect on recommending a product.

Model 2 is a two stage Heckman probit regression and the results are shown in Table 4. The first stage estimates the effect of individual and product estimates on review propensity. We find that mono purists, poly purists and mono mixers are more likely to review products compared to poly mixers

supporting hypothesis 1a. We find that individuals who own more games are proportionally more likely to review compared to individuals who own fewer games (β=0.0371) supporting hypothesis 2a. We also find that individuals who have more friends on the platform are more likely to review products compared to individuals who have fewer friends on the platform (β=0.127) supporting hypothesis 3a. Playtime has a strong and significant effect on review propensity. Individuals who play a game for long periods of time are much more likely to review that game (β=0.253). We also find that more typical games and higher priced games are more likely to be reviewed.

The second stage of the Heckman Selection model estimates the effect of individual and product characteristics on recommendations after accounting for the review propensity. In Table 3, when we study the effect of different segments of consumers on ratings, we find that mono purists, poly purists and mono mixers do not differ from poly mixers on their ratings. However, after controlling for review propensity we find that mono purists, poly purists and mono mixers are significantly more likely to provide higher ratings compared to poly mixers thus supporting hypothesis 1b. After controlling for the propensity to review, number of friends by an individual continues to have a significant positive effect on ratings (β=0.142). Individuals who have more friends on the platform are more likely to recommend products positively which supports hypothesis 2b. We find that after controlling for the propensity to review, the negative effect of the number of games owned by an individual on ratings (as seen in Table 3) vanishes and hence hypothesis 3b is not supported. Individuals who play the game for longer periods of time are more likely to give higher ratings supporting hypothesis 4b.

Comparing estimates of Table 4 with estimates of Table 3, we find that there is a significant difference in the estimates of all the variables between the two tables. This indicates that not accounting for review propensity while studying online ratings can result in biased findings.

The third model estimates the effect of individual and product characteristics on propensity to post positive and negative reviews. We find that individuals who own more games are more likely to post more negative and also more likely to post more positive reviews compared to individuals who own fewer games. However, we find that the estimate for propensity to post negative reviews (β=0.255) is approximately four

times greater compared to the estimate for propensity to post positive reviews (β=0.0617). This indicates that individuals who own more games are more likely to review games negatively than they are to review games positively. Individuals who have more friends on the platform are more likely to post more negative and positive reviews compared to individuals who have fewer friends. The estimate for the propensity to post positive reviews (β=0.345) is approximately three times the estimate for the propensity to post negative reviews (β=0.123) indicating that these individuals are more likely to post positive reviews than post negative reviews.

We also find that overall playtime of a game by an individual has a significant effect on an individuals' propensity to rate games positively. Individuals are much more likely to rate games positively that they play for long periods of time (β=0.710). Surprisingly, we find that overall playtime also has a significant effect on negative reviews. Individuals are also more likely to review games that they play for longer periods of time negatively (β=0.203). One possible reason could be due to changes or updates to the game. If a developers updates games with new features that users who are heavily invested in the game do not like, then these users are much more likely to review the game negatively. Higher priced games and games which have higher sales are more likely to be positively reviewed and less likely to be negatively reviewed. We also find that mono purists, poly purists and mono mixers are more likely to write more positive reviews compared to poly mixers. We find that compared to poly mixers, mono mixers are more likely to write more negative reviews.

In Appendix 1, as robustness tests, I use continuous measures of variety seeking and atypicality seeking and we find that interaction term of variety seeking and atypicality seeking is negative and significant (β=-1.323), suggesting that individuals with high variety seeking and atypicality seeking tendencies (poly mixers) are less likely to review products compared to the other three groups.

9. Discussion

Prior studies that have used ratings from online reviews to study the effect of consumer heterogeneity and product category structure on product appeal studies do not incorporate biases caused

due to review propensity as they do not capture product purchase or ownership data. This has been mostly due to lack of availability of ownership data. Product reviews and ratings from major sites are widely available and can be scraped easily but majority of the sites either do not have product purchase data or do not provide them and hence prior studies have not accounted for review propensities while studying online ratings which could lead to biased findings. In this study, I use both product ownership and review data and hence incorporate review propensity to study online ratings. I draw upon research on category structure and consumer heterogeneity from sociology and integrate this with literature on online reviews biases from information systems and marketing.

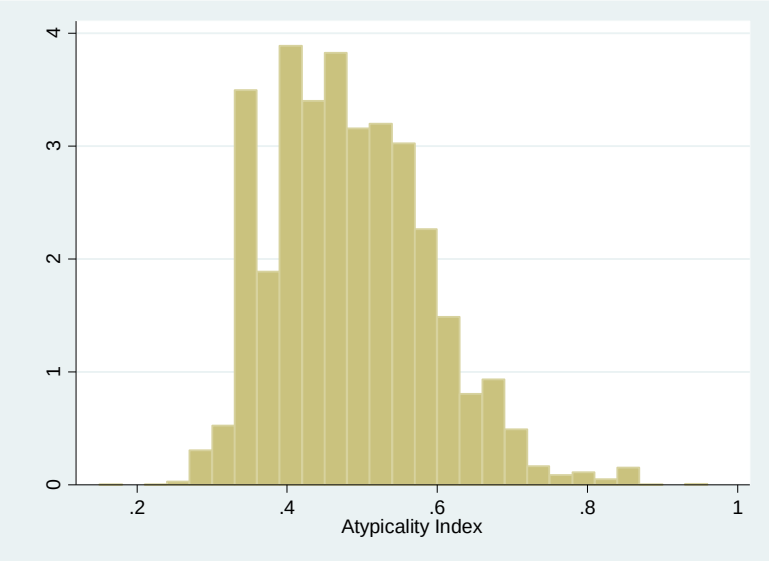
This study provides multiple avenues for extension for future studies. Average ratings of products generally decline over time (Duan et al. 2008; Li and Hitt 2008, Zhu and Zhang 2010) and this has been attributed to early reviewers being more excited about the product. Future researchers can study if there is a difference in change of ratings over time for atypical products compared to more typical products.

Unfortunately, I do not have demographic information about individuals such as age, sex, education and such. If demographic information of consumers can be collected, then it would be interesting to incorporate demographic information and try to understand how different consumer segments differ in their tastes for atypical products.

This research also has several practical implications for platform owners and review sites. First, few consumers make an effort to write online reviews. I find that on Steam only one in five consumers opt in to write a review. Second, even among the set of consumers who review, majority of the consumers only review very few products even though they purchase many more products. Consumers do not trust online reviews as they believe that sites contain biased reviews. Eliminating or minimizing bias in online reviews is very critical for many product sites as it would improve trust in the online ratings which can increase consumer traffic to the sites.

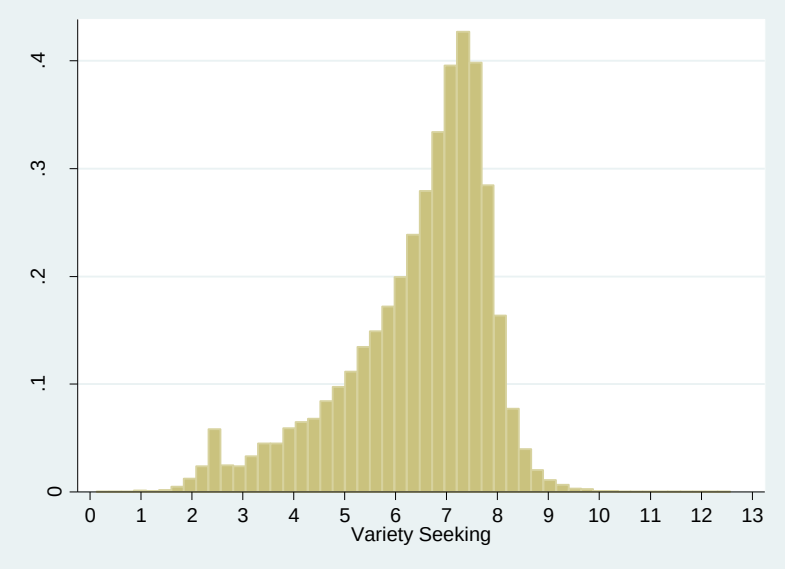
To improve trust in online reviews, platform owners would want more people to participate and ensure that only genuine consumers write unbiased reviews. To ensure even representation, platform owners can specifically target unrepresented consumer segments to participate in online reviews. Platform owners can also delink ratings from reviews and ask consumers to provide ratings instead of asking consumers to write detailed reviews as well as provide ratings. Another method of reducing bias and eliminating fake reviews is to select only a random set of consumers for each product and invite them to provide ratings and reviews.

Figure 1: Histogram of Atypicality Index



N=3,985

Figure 2: Histogram of Variety Seeking Index



N=130,988

Table 1: Descriptive Statistics

Variable	# Obs.	Mean	Std. Dev.	Min	Max
Individual Level Variables					
Count of Games Owned$_i$	130,988	63.06	111.28	0.00	6125.00
Variety Seeking$_i$	130,988	6.41	1.45	0.13	12.56
Atypicality Seeking$_i$	130,988	0.47	0.05	0.26	0.87
Count of Friends$_i$	130,988	26.17	41.71	0.00	923.00
Days Since Join from 2017$_i$	130,988	2136.40	193.11	1842.70	2533.20
Average Playtime of Individual$_i$	130,988	1895.49	2287.41	0.27	43899.60
Poly Mixer$_i$	130,988	0.37	0.48	0.00	1.00
Poly Purist$_i$	130,988	0.13	0.33	0.00	1.00
Mono Mixer$_i$	130,988	0.13	0.34	0.00	1.00
Mono Purist$_i$	130,988	0.37	0.48	0.00	1.00
Product Level Variables					
Price of Game$_j$	3,985	11.14	11.73	0.00	399.00
Game Sales$_j$	3,985	3285.66	11087.92	1.00	274102.00
Days Since Release$_j$	3,985	1020.32	809.00	17.00	6618.00
Average Playtime of Game$_j$	3,985	361.59	960.02	0.02	30722.69
Atypicality of Game$_j$	3,985	0.49	0.11	0.16	0.95
Review and Playtime Variables					
Recommended$_{ij}$	30,530	0.87	0.34	0.00	1.00
Days Since Posting$_{ij}$	30,530	655.82	389.59	0.00	2235.00
Reviewed$_{ij}$	2,461,219	0.01	0.11	0.00	1.00
Playtime Forever$_{ij}$	2,461,219	2330.57	7136.07	21.00	99997.00

Table 2: Correlations

1	1															
2	.	1														
3	.	0.11*	1													
4	0.11*	0.08*	0.03*	1												
5	0.02*	-0.01	0.06*	-0.07*	1											
6	-0.01*	-0.01	0.03*	-0.04*	0.29*	1										
7	0.03*	0.04*	0.03*	0.05*	0.13*	-0.09*	1									
8	-0.01*	-0.02*	0.01*	0.01*	0.02*	-0.01*	0.13*	1								
9	-0.01*	-0.02*	-0.01*	-0.03*	0.18*	0.59*	-0.12*	-0.02*	1							
10	-0.01*	0	0.01	-0.04*	0.30*	0.67*	-0.08*	-0.01*	0.65*	1						
11	0.00*	0.01	0.03*	0.01*	-0.08*	-0.01*	0.02*	0.00*	-0.39*	-0.53*	1					
12	0	-0.02*	-0.04*	0.01*	-0.14*	-0.37*	0.01*	-0.01*	0.03*	-0.43*	-0.10*	1				
13	0.00*	0.01*	-0.01*	0.04*	-0.24*	-0.64*	0.08*	0.02*	-0.57*	-0.55*	-0.13*	-0.11*	1			
14	0.02*	-0.04*	-0.13*	0.07*	0.00*	0.12*	-0.03*	0.03*	0.09*	0.09*	-0.01*	-0.04*	-0.08*	1		
15	0.02*	0.08*	0.13*	0.25*	-0.12*	-0.13*	0.05*	-0.01*	-0.15*	-0.12*	0.05*	0.02*	0.11*	-0.14*	1	
16	-0.03*	0.10*	0.30*	0.03*	-0.05*	-0.11*	0.01*	0.05*	-0.12*	-0.09*	0.04*	0.00*	0.10*	-0.22*	0.35*	1
17	-0.01*	-0.01*	-0.12*	-0.06*	0.05*	0.17*	-0.03*	0.00*	0.28*	0.18*	-0.11*	0.01*	-0.16*	0.10*	-0.31*	-0.26*

1 Review

2 Recommend

3 Days Since Posting

4 Playtime Forever

5 # Games Owned

6 Variety Seeking

7 Count of Friends

8 Days Since Join

9 Atypicality Seeking

10 Poly Mixer

11 Poly Purist

12 Mono Mixer

13 Mono Purist

14 Price of Game

15 Sales of Game

16 Days Since Release

17 Atypicality of Game

Table 3 Effect of Product and Individual characteristics on Recommendation

	(1) Recommend	(2) Recommend	(3) Recommend
Count of Games Owned$_i$	-0.109***	-0.141***	-0.140***
	(0.014)	(0.018)	(0.019)
Count of Friends$_i$	0.0932***	0.0972***	0.0966***
	(0.010)	(0.012)	(0.012)
Days since Join (1000's)$_i$	-0.104*	-0.141**	-0.143**
	(0.054)	(0.068)	(0.068)
Average Playtime of Individual$_i$	-0.102***	-0.273***	-0.273***
	(0.016)	(0.021)	(0.021)
Overall Playtime$_{ij}$		0.244***	0.244***
		(0.008)	(0.008)
Average Playtime of Game$_j$		-0.181***	-0.181***
		(0.011)	(0.011)
Atypicality of Game$_j$		0.800***	0.813***
		(0.107)	(0.108)
Game Sales$_j$		0.0696***	0.0691***
		(0.009)	(0.009)
Price of Game$_j$		0.133***	0.133***
		(0.009)	(0.010)
Days since Release$_j$ (1000s)		0.0912***	0.0909***
		(0.014)	(0.014)
Days since Posting of Review$_{ij}$ (1000s)		0.397***	0.397***
		(0.000)	(0.000)
Mono Purist$_i$			0.0175
			(0.039)
Poly Purist$_i$			0.00625
			(0.038)
Mono Mixer$_i$			-0.0435
			(0.043)
Constant	2.282***	1.642***	1.646***
	(0.195)	(0.265)	(0.266)
Observations	56038	32673	32673
Log Likelihood	-22135.6	-10461.2	-10459.9

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4 Two Stage Heckman Selection on Review Propensity and Recommendation

	(1) Recommend	(2) Recommend	(3) Recommend
2nd Stage Recommend			
Count of Games Owned$_i$	0.0572***	-0.00528	0.00302
	(0.007)	(0.013)	(0.012)
Count of Friends$_i$	0.113***	0.144***	0.142***
	(0.004)	(0.007)	(0.007)
Days since Join (1000's)$_i$	-0.247***	-0.258***	-0.256***
	(0.024)	(0.031)	(0.031)
Average Playtime of Individual$_i$	0.102***	-0.110***	-0.108***
	(0.007)	(0.019)	(0.018)
Overall Playtime$_{ij}$		0.297***	0.296***
		(0.011)	(0.010)
Average Playtime of Game$_j$		-0.123***	-0.121***
		(0.011)	(0.010)
Atypicality of Game$_j$		0.117*	0.140**
		(0.063)	(0.060)
Game Sales$_j$		0.0180***	0.0180***
		(0.006)	(0.005)
Price of the Game$_j$		0.0482***	0.0493***
		(0.008)	(0.008)
Days since Release$_j$ (1000s)		-0.0950***	-0.0966***
		(0.010)	(0.009)
Days since Posting of Review$_{ij}$ (1000s)		0.0967***	0.0938***
		(0.000)	(0.000)
Mono Purist$_i$			0.0449***
			(0.017)
Poly Purist$_i$			0.0389**
			(0.017)
Mono Mixer$_i$			0.0436*
			(0.023)
Constant	-3.144***	-2.637***	-2.723***
	(0.082)	(0.165)	(0.156)

1st Stage – Review	Review	Review	Review
Count of Games Owned$_i$	0.0696***	0.0294***	0.0371***
	(0.006)	(0.007)	(0.008)
Count of Friends$_i$	0.105***	0.129***	0.127***
	(0.004)	(0.005)	(0.005)
Days since Join (1000's)$_i$	-0.242***	-0.237***	-0.235***
	(0.025)	(0.029)	(0.029)
Average Playtime of Individual$_i$	0.114***	-0.0474***	-0.0472***
	(0.007)	(0.008)	(0.008)
Overall Playtime$_{ij}$		0.253***	0.253***
		(0.002)	(0.002)
Average Playtime of Game$_j$		-0.0838***	-0.0834***
		(0.003)	(0.003)
Atypicality of Game$_j$		-0.0718***	-0.0434*
		(0.026)	(0.026)
Game Sales$_j$		0.000266	0.000956
		(0.003)	(0.003)
Price of Game$_j$		0.0170***	0.0193***
		(0.003)	(0.003)
Days since Release$_j$ (1000s)		-0.122***	-0.122***
		(0.004)	(0.004)
Mono Purist$_i$			0.0443***
			(0.016)
Poly Purist$_i$			0.0391**
			(0.016)
Mono Mixer$_i$			0.0583***
			(0.022)
Constant	-3.223***	-2.838***	-2.917***
	(0.082)	(0.098)	(0.099)
athrho			
Constant	4.187***	2.415***	2.460***
	(0.397)	(0.366)	(0.347)
Observations	4348001	2494753	2494753
Log Likelihood	-315654.8	-166779.6	-166736.5

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5 Effect of Product and Individual characteristics on Review Propensity on Positive and Negative Reviews

Negative Review			
Count of Games Owned$_i$	0.343***	0.236***	0.255***
	(0.027)	(0.034)	(0.036)
Count of Friends$_i$	0.122***	0.127***	0.123***
	(0.020)	(0.024)	(0.024)
Days since Join (1000's)$_i$	-0.463***	-0.334**	-0.322**
	(0.115)	(0.141)	(0.141)
Average Playtime of Individual$_i$	0.468***	0.390***	0.389***
	(0.033)	(0.042)	(0.042)
Overall Playtime$_{ij}$		0.205***	0.203***
		(0.014)	(0.014)
Average Playtime of Game$_j$		0.0641***	0.0652***
		(0.021)	(0.021)
Atypicality of Game$_j$		-1.191***	-1.141***
		(0.175)	(0.180)
Game Sales$_j$		-0.154***	-0.151***
		(0.016)	(0.017)
Price of Game$_j$		-0.137***	-0.130***
		(0.015)	(0.015)
Days since Release$_j$ (1000s)		-0.555***	-0.556***
		(0.030)	(0.030)
Mono Purist$_i$			0.0824
			(0.078)
Poly Purist$_i$			0.117
			(0.079)
Mono Mixer$_i$			0.246**
			(0.096)
Constant	-10.82***	-8.707***	-8.928***
	(0.417)	(0.541)	(0.551)
Positive Review			
Count of Games Owned$_i$	0.152***	0.0444**	0.0617***
	(0.017)	(0.018)	(0.020)
Count of Friends$_i$	0.299***	0.351***	0.345***
	(0.012)	(0.012)	(0.012)

Days since Join (1000's)$_i$	-0.654***	-0.614***	-0.609***
	(0.066)	(0.071)	(0.071)
Average Playtime of Individual$_i$	0.274***	-0.228***	-0.227***
	(0.019)	(0.021)	(0.021)
Overall Playtime$_{ij}$		0.711***	0.710***
		(0.006)	(0.006)
Average Playtime of Game$_j$		-0.271***	-0.270***
		(0.009)	(0.009)
Atypicality of Game$_j$		-0.0320	0.0393
		(0.069)	(0.069)
Game Sales$_j$		0.0392***	0.0409***
		(0.009)	(0.009)
Price of the Game$_j$		0.0956***	0.100***
		(0.006)	(0.006)
Days since Release$_j$ (1000s)		-0.293***	-0.295***
		(0.010)	(0.010)
Mono Purist$_i$			0.106***
			(0.039)
Poly Purist$_i$			0.0845**
			(0.040)
Mono Mixer$_i$			0.113**
			(0.054)
Constant	-6.805***	-5.886***	-6.066***
	(0.225)	(0.250)	(0.251)
Observations	4348001	2494753	2494753
Log Likelihood	-315666.6	-166350.7	-166312.1

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

References

Bagozzi, R. P., & Dholakia, U. M. 2006. Open source software user communities: A study of participation in Linux user groups. *Management science*, *52*(7), 1099-1115.

Chevalier, J. A., and Mayzlin, D. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research* (43:3), pp. 345-354.

DiMaggio, Paul, 1987. Classification in art. *American Sociological Review* 52, 440-455.

Driver, Michael J. and Siegfried Streufert. 1964. "The General Incongruity Adaptation Level' (GIAL) Hypothesis: An Analysis and Integration of Cognitive Approaches to Motivation." Paper No. 114. Institute for Research in the Behavioral, Economic and Management Sciences. Krannet Graduate School of Management, Purdue University. West Lafayette, IN.

Duan, W., Gu, B., Whinston, A. B. 2008. Do online reviews matter? — An empirical investigation of panel data. *Decision Support Systems*, 45 (4), 1007–1016.

Elberse, A., 2008. Should you invest in the long tail?. *Harvard business review*, *86*(7/8), p.88.

Feiler, D. C., & Kleinbaum, A. M. 2015. Popularity, Similarity, and the Network Extraversion Bias. *Psychological Science (Sage Publications Inc.), 26*(5), 593-603. doi:10.1177/0956797615569580

Gardenfors, Peter. 2004. *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: The MIT Press.

Givon, M. 1984. Variety-seeking through brand switching. *Marketing Science*, *3*,1–22.

Goldberg, A., Hannan, M.T. and Kovács, B., 2016. What does it mean to span cultural boundaries? Variety and atypicality in cultural consumption. *American Sociological Review*, *81*(2), pp.215-241.

Hannan, Michael T., Laszlo Polos, and Glenn R. Carroll. 2007. *Logics of Organization Theory: Audiences, Codes, and Ecologies*. Princeton, NJ: Princeton University Press.

Hsu, Greta. 2006. "Jacks of All Trades and Masters of None: Audiences' Reactions to Spanning Genres in Feature Film Production." *Administrative Science Quarterly* 51:420–450.

Hu, N., Pavlou, P. A., Zhang, J. 2009. Overcoming the J-shaped Distribution of Product Reviews. *Communications of the ACM*, 52 (10), 144–147.

Hu, Y., Li, X. 2011. Context-Dependent Product Evaluations: An Empirical Analysis of Internet Book Reviews. *Journal of Interactive Marketing*, 25 (3), 123–133.

Josh Lerner, and Jean Tirole 2001, "The Open Source Movement: Key Research Questions", *European Economic Review*, Elsevier, vol. 45, n. 4-6, pp. 819–826.

Koh, N. S., Hu, N., Clemons, E. K. 2010. Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, 9 (5), 374–385.

Kovács, B., & Johnson, R. 2014. Contrasting alternative explanations for the consequences of category spanning: A study of restaurant reviews and menus in San Francisco. *Strategic Organization*, 12: 7–37.

Kovács, Balázs, and Michael T. Hannan. 2010. "The Consequences of Category Spanning Depend on Contrast." *Research in the Sociology of Organizations* 31:175–201.

Kovács, Balázs, and Michael T. Hannan. 2015. "Conceptual Spaces and the Consequences of Category Spanning." *Sociological Science* 2: 252-286.

Li, X., Hitt, L. M. 2008. Self-Selection and Information Role of Online Product Reviews. *Information Systems Research*, 19 (4), 456–474.

Liu, Y. 2006. "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing* (70:3), pp. 74-89.

Markides, C. C. and Williamson, P. J. 1994 "Related Diversification, Core Competences and Corporate McAlister, L., & Pessemier, E. (1982).Variety- seeking behavior: An interdisciplinary review. *Journal of Consumer Research, 9*, 311–322.

Ming D. Leung and Weiyi Ng. 2014. "An idiosyncrasy credit or a generalist discount? Conditional advantages to working broadly in a virtual labor market". IRLE Working Paper No. 103-14.

Mudambi, S. M. and Schuff, D. 2010. "What Makes a Helpful Review? A Study of Customer Reviews on Amazon.com," *MIS Quarterly* (34:1), pp.185-200.

Negro, G., Hannan, M. T., & Rao, H. 2010. Categorical contrast and audience appeal: Niche width and critical success in winemaking. *Industrial and Corporate Change*, 19: 1397–1425.

Ollivier, Michele. 2008. "Modes of Openness to Cultural Diversity: Humanist, Populist, Practical, and Indifferent." *Poetics* 36(1):120–47.

Peterson, Richard A. 1992. "Understanding Audience Segmentation: From Elite and Mass to Omnivore and Univore." *Poetics* 21(4):243–58.

Peterson, Richard A. 2005. "Problems in Comparative Research: The Example of Omnivorousness." *Poetics* 33(5–6):257–82.

Relish, Michael, 1997. It's not all education: network measures as sources of cultural competency. *Poetics* 25, 121–139.

Trenz, M., & Berger, B. 2013. Analyzing Online Customer Reviews-An Interdisciplinary Literature Review And Research Agenda. In *ECIS* (p. 83).

Venkatesan. M. 1973. "Cognitive Consistency and Novelty Seeking." in *Consumer Behavior: Theoretical Sources,* eds.Scott Ward and Thomas S. Robenson. Englewood Cliffs. NJ: Prentice-Hall, 355-384.

Widdows, Dominic. 2004. *Geometry and Meaning*. Stanford, CA: CSLI Publications.

Yin, D., Bond, S., and Zhang, H. 2014. "Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews," *MIS Quarterly* (38:2), pp.539-560.

Zhang, Xiaoquan (Michael) and Feng Zhu. 2011. "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia." *American Economic Review*, 101(4):1601-15.

Zhu, F., Zhang, X. (Michael) 2010. Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. *Journal of Marketing*, 74 (2), 133–148.

Zuckerman, E. W. 1999. The categorical imperative: Securities analysts and the illegitimacy discount. *American Journal of Sociology*, 104: 1398–1438.

Appendix

Appendix 1: Effect of Variety Seeking and Atypicality Seeking on Review propensity

	(1) Review	(2) Review
Count of Games Owned$_i$	0.0725***	0.0698***
	(0.020)	(0.020)
Count of Friends$_i$	0.320***	0.319***
	(0.012)	(0.012)
Days since Join (1000's)$_i$	-0.582***	-0.572***
	(0.071)	(0.071)
Average Playtime of Individual$_i$	-0.145***	-0.143***
	(0.021)	(0.021)
Overall Playtime$_{ij}$	0.643***	0.643***
	(0.006)	(0.006)
Average Playtime of Game$_j$	-0.221***	-0.219***
	(0.009)	(0.009)
Atypicality of Game$_j$	-0.142**	-0.129**
	(0.063)	(0.063)
Game Sales$_j$	0.0118	0.0110
	(0.008)	(0.008)
Price of the Game$_j$	0.0612***	0.0633***
	(0.006)	(0.006)
Days since Release$_j$ (1000s)	-0.317***	-0.312***
	(0.010)	(0.010)
Variety Seeking$_i$	0.0115	0.632***
	(0.018)	(0.118)
Atypicality Seeking$_i$	-1.251**	7.672***
	(0.532)	(1.674)
Variety Seeking$_i$ * Atypicality Seeking$_i$		-1.323***
		(0.254)
Constant	-5.350***	-9.541***
	(0.319)	(0.810)
Observations	2494753	2494753
Log Likelihood	-155979.0	-155946.7

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$