Peer influence of production and consumption behaviour in an online social network of collective learning

Abhishek Samantray, Massimo Riccaboni

Email abhishek.samantray@imtlucca.it, massimo.riccaboni@imtlucca.it
Address Laboratory for the Analysis of Complex Economic Systems (AXES), IMT School for Advanced Studies Lucca, Italy

To cite, use:-
Abhishek Samantray, Massimo Riccaboni (2020)
Peer influence of production and consumption behaviour in an online social network of collective learning.
Online Social Networks and Media, Vol. 18, Pg. 100088
doi: https://doi.org/10.1016/j.osnem.2020.100088

Abstract We study peer influence of production and consumption of projects in the Scratch community, an online platform developed by MIT Media Lab and targeted for young children, where users collectively learn programming by creating and sharing projects. We investigate if Scratchers are influenced by the popularity of their peers' projects and their peers' preferences for consuming from specific baskets of projects.

We find that the popularity of Scratchers' projects is significantly influenced by the production popularity of their peers. Testing for heterogeneity in influence, we find that Scratchers are not influenced by specific peers who might have highly popular projects, instead it seems that they are influenced by just the aggregate popularity of all peers. We find that Scratchers who have a minimum activity of one month on the platform are more susceptible to peer influence. Scratchers with high tendency to create projects by rebuilding on existing projects on the platform tend to have significant improvements in their future production popularity (due to influence from peers' production popularity) only in the short term and not in the long run. We also disentangle a self decision making mechanism from other mechanisms that might explain the channel of influence: we find that a significant proportion of the estimated influence from peers is mediated via Scratchers' decision to create new projects. This highlights Scratchers' subsequent behavioural decisions in response to existing popularity of peers' projects.

We find evidence of polarized consumption patterns on the platform, i.e., there are certain groups of projects (discovered in an unsupervised manner based on co-consumption patterns) for which Scratchers have high specificity. We do not make claims about how such groups form on the platform - for example, whether it is a conscious choice or is a result of the way the platform is organized. However, we find that such polarization is not a consequence of Scratchers being influenced by their peers' consumption patterns.

Keywords peer influence, causal mediation mechanism, social influence, homophily, online social network, human behaviour, computational social science

Significance: Today various online educational platforms facilitate collective ways of learning. The literature on peer influence on educational outcomes presents mixed evidences, i.e., both positive and negative influences. Since learning via educational platforms is gaining increasing interest, it is important for business and economic policy designers to know the real impact of peers activities on the choices and educational outcomes of users. On the Scratch platform which is targeted for young children to learn programming, we find that users are significantly influenced by the production popularity of their peers, but not by their peers' consumption patterns. We contemplate that knowledge of such behavioural nature would be useful to design platforms where the collective educational outcome is maximized.

Contents

1 Introduction

Education is a vital tool of empowerment. Peers' behaviour can play an important role in various aspects of one's education process [1]. Studying peer influence in learning environments can therefore be helpful for various business, economic, and government policy designs. How co-learners influence educational and social outcomes has been studied extensively in physical contexts like schools and universities [2, 3]. With the advent of various online education media and many users joining such sites, it is important to study peer influence in digital platforms as well. Such platforms usually encourage learning by various forms of collective interactions such as discussions in forums, building collaborative projects, private communications, and others [4]. A possibility of peer influence arises since users are usually aware of others' shared activities. In this study, we investigate peer influence in the Scratch platform which has a structure of learning through collective activities. Scratch, made public in 2007, is an online community designed by the MIT Media Lab for young people to learn programming. Scratchers produce and share visual projects built using programming codes. Scratchers also consume others' projects in various ways which include viewing, commenting, loving, downloading, etc. Scratchers can know about the activities of their peers, other users whom they "follow" on the platform, via activity feeds and also by manual visits to their project pages. This creates a potential channel of influence on various behaviours. In the first five years of Scratch's public activities [5], during which about 1 million users joined the platform and about 2 million projects were created, we investigate peer influence on two behaviours – one relates to production of projects, and the other relates to consumption of projects.

Understanding of peer influence estimation has been shaped by contributions from academics and practitioners in various fields including marketing, sociology, and economics. To infer peer influence, the most ideal situation would be to impute peers' behaviour at random and measure its average effect on Scratchers' behaviour. Marketing scientists have used such behavioural imputations in various online platforms to measure peer influence [6]. However such experimental situations are usually not feasible, especially in non-artificial circumstances, for several reasons including ethics and permissions to perform such experiments [7, 8, 9]. In non-experimental settings, obtaining unbiased estimates of peer in-

fluence is a challenging task because both individuals and their peers can affect each others' behaviour (reflection problem [10], and so observed clustering of behaviour in networks is often a result of the following effects: own tendency for the behaviour, peers' influence on behaviour, and exogenous and endogenous network formation processes (homophily, selection, reciprocity, etc.) leading to observed peers' behaviour. Dynamic observations help to separate changes in individual behaviour due to peers' influence from effects arising due to alternative mechanisms. Sociologists have used agent-based models [11, 12, 13, 14, 15] to explain the coevolution of network and behaviour. However, this method requires agents to have full knowledge of the network, which is rarely the case when the network is populated by a large number of agents. Economists have used estimation strategies that usually require strong assumptions [16, 17] and are specific to the structural models employed [18]. Sometimes exogenous component of peer influence (arising from past) is not estimated separately from the contemporaneous effect arising due to simultaneous determination of network formation and behavioural influence [19].

We employ a quasi-experimental method to identify peer influence. We assume Scratchers have a Markov nature of decision making, i.e., their future decisions (e.g., whom to follow, what to produce and consume) are driven only by the current state of activities on the platform. Conditional on all activities upto a given time t, we estimate peer influence at t on a future time $t+j$ as the effect of peers' behavioural state at t on Scratchers' subsequent change in behaviour upto $t+j$. The quasi-experiment consists of observations at two time periods – t, $t+j$ – and treatment status is assigned at t based on intensity of peers' behaviour (high or low) at t[1]. The treated group has Scratchers whose peers have high degree of behaviour under study. To conceptualize the treatment status as a random assignment, the control group is adjusted by matching exactly on personal and peers' characteristics of Scratchers in the treated group such that all confounding factors are balanced across the two groups. Below are the main results and contributions of our investigation:

1. The definition of treatment follows directly from the Markovian nature of decision making by Scratchers. The treatment, peers' behavioural state at t, is a measure that summarizes peers' behaviour upto t. It captures only the cumulative information of peers' behaviour upto t and neglects the historical pattern of its evolution.

– If peers' projects are popular (as measured by accumulated 'loves' on the projects) at t, the popularity of Scratchers' projects increases in future periods. This effect is persistent and is S-shaped along the axis representing future periods. A minimum engagement of about a month on the platform makes Scratchers more susceptible to such influence, however higher engagement does not necessarily increase the susceptibility. Remixing is a key property of the Scratch platform – users can build projects on top of existing projects by modifying or introducing new elements. Developers, users whose projects tend to be mostly remixed and not new projects, are influenced more in the short term and free-style producers, Scratchers who create new and remixed projects in same proportion, tend to be influenced in later periods only and not in the immediate period.

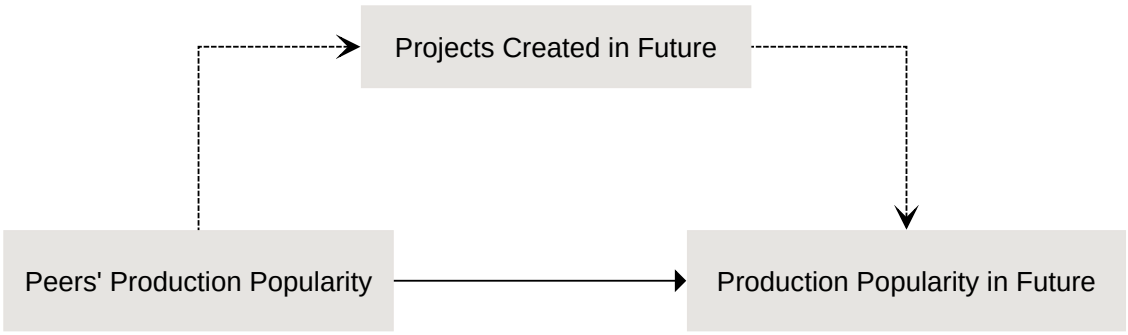
– As shown in Figure 1, we investigate if the observed peer influence in production popularity is caused due to production-related decisions made by Scratchers. 40-50 percent of the total effect of peers' production quality on Scratchers' future production popularity is mediated via their creation of new projects in future. This channel emphasizes the role of decision-making under influence of peers' behaviour.

– Scratchers are not influenced by their peers' consumption patterns – if peers tend to 'love' projects from a specific community of projects, it does not influence Scratchers to develop a similar preference in future.

– From a methodological perspective, we improvise on the approach by Aral et al. [20] in two ways: First, we use exact matching to obtain a control group where Scratchers are similar to those in the treated group except for the peers' behaviour under study; this helps to minimize bias to a large extent, compared to using propensity score matching [20]. Second, by ensuring balance of peers' characteristics (in addition to individual characteristics) we control not only for homophily [20], but also for other confounding effects including selection and endogenous processes involved in network formation [13], and own behavioural tendencies.

In the next section (Section 2), we describe the Scratch platform and the data we analyze in more details. Then, in Section 3, we provide descriptions of produc-

Figure 1: Peer Influence Mechanism



Popularity of peers' projects affects the popularity of Scratchers' projects in future (solid path). 40-50% of such peer influence on production popularity is mediated via Scratchers' creation of new projects in future (dotted path).

tion and consumption behaviour that are helpful to gain insight into aggregate behavioural patterns over the first five years, and these descriptions are used later to augment the results on peer influence. This is followed by, in Section 4, a detailed description of the methods we use to identify peer influence and the results of peer influence analysis. Since we observed a significant peer influence for production behaviour, we investigate in the next section (Section 5) how this happens, and particularly if Scratchers make any personal decisions leading to their improved outcomes in future periods. Finally, in Section 6, we make a discussion based on our findings, and provide suggestions for further research.

2 Data: Scratch Community

We analyze users' behaviour in the Scratch community[2], an online educational platform created and maintained by the Lifelong Kindergarten Group at MIT Media Lab. Users come from various countries. The platform, designed for children in schools, serves as an educational media to collectively learn programming by creating and sharing interactive objects. An interactive object created on the platform is called a *project*, which is usually an animation, game, or simulation created using the Scratch programming language (SPL) [21]. Projects are composed from animated objects called *sprites*. SPL employs drag-and-drop programming method to create projects, using Scratch Authoring Environment (SAE), by assembling basic visual elements called *blocks*. The online platform was created in March 2007. SPL has had two major development versions - Scratch 1.x (1.0 to 1.4) and Scratch 2.0 (released in May 2013). To build or edit a project in 1.x versions, users had to download the Scratch editor software (offline version) to access the SAE. Users could then (optionally) share the projects in the online community. In version 2.0, which replaced 1.x, users can access SAE both online and offline.

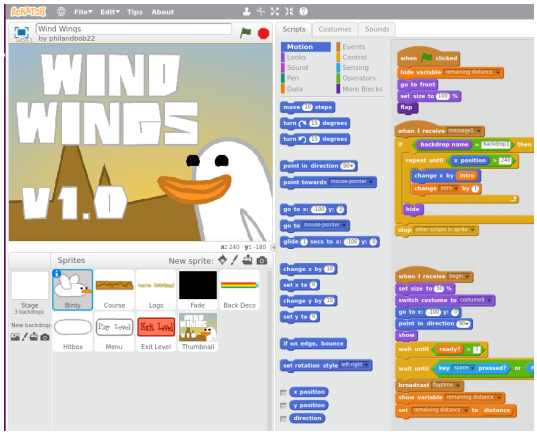
Data from March 2007 to March 2012 was provided by the MIT Media Lab under the Scratch Research Data Sharing Agreement [5]. It consists of various metadata, corresponding to the descriptions below, of all users and their friendship formations, and of all the projects created during this period. Hence this forms a complete data set of time-stamped users' friendship network and production and consumption of projects for the first 5 years.

In the Scratch community users can (i) produce projects, (ii) consume projects, (iii) follow other users as friends, and (iv) create and comment on galleries. Such collective action in Scratch community is analogous to activities in the social media platform Facebook where contents (posts or status updates) are produced and consumed by the platform users, and users can also follow each other. Projects created (Fig.2a) on the Scratch platform can be of two types - *new*, and *remix*. A new project, as is suggestive, is a fresh project created by a user and shared on the website. A remix project shared by a user is a project that is created by modifying an already existing project (new/remix) on the platform. After a project

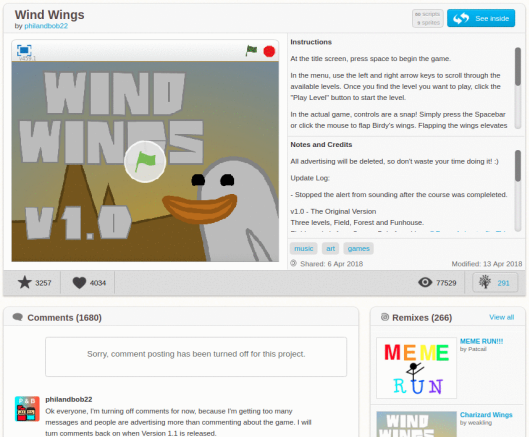
2. https://scratch.mit.edu

Figure 2: Projects: Production & Consumption

(a) Production Perspective



(b) Consumption Perspective



Production and consumption perspectives of a project in Scratch 2.0. (a) A project is created by composing various sprites that have codes, images, and sound associated to them. (b) A shared project is available to other users for viewing, downloading, loving, and commenting.

is shared by a user, it can be consumed (Fig.2b) by other users on the platform. Consumption of a project on the Scratch website refers to the following interactions with the project by logged-in users: *viewing*, *downloading*, *loving*, *commenting*, and *favoriting*. Each form of consumption of a project by a user is recorded only once - the first time the user interacts with it. Views, downloads, and loves of a project are anonymous records, i.e, the names of the users who interacted with the project by such forms are not recorded. Friendships represent unidirectional relationships between users. A user can choose to follow any other user on the platform. Once logged-in, a user can see the latest projects of the users he is following in a dedicated section. Users can also create *galleries* which are collections of projects. Users can view and comment on galleries. Projects and galleries can also be *tagged* by their creators. Tag names are not pre-defined on the platform, and new tag names are created when users tag projects and galleries with non-existing tag names. Projects and galleries can have common tags names.

Additionally, selected projects are displayed in the *front page* of the Scratch website due to various criteria (most remixed, most viewd, etc.). This selection is automated. Within each category, the three most recently added projects are displayed at any given point. There is a section on the front page for *featured projects* (three projects at one time), projects in this section are manually added by users who are Scratch website administrators based on popularity and appeal of projects. For galleries, there are two sections on the front page, one is a section called *featured galleries*, and the other is called *studio design*. Addition of galleries to these sections are controlled by administrators. A user can at some point be assigned as a *curator* by administrator. The curator selects projects for the Scratch website's front page section labeled 'Curated By'. This section displays three recent projects selected by the curator. There is only one active curator at a time.

The dedicated section where Scratchers can see the activities of their peers in real-time is called *'What's Happening?'* [22]. Here Scratchers can see the following recent activities of their peers – sharing (creation) of projects, remixing, love-its,, favorites, following (users, studios). This is an important channel of information about peers' activities; if a Scratcher is following many others, he would most likely be influenced by activities of those which appear frequently via this feed. It is important to note that a Scratcher can know which projects his peers are favoriting via activity feed, however the projects which receive the favorite clicks do not show such counts on the project page. We see in Figure 2(b) that favorites (star symbol) count are visible, however this is for the latest version of Scratch. During the period 2007-12 for which data is available, SPL versions 1.x were in place, and favorites count was not visible on the project page. The love-it counts (and all other forms of consumption except favorites) on the other hand are shown on the project pages, and is public information; this forms the difference between favorites and love-its.

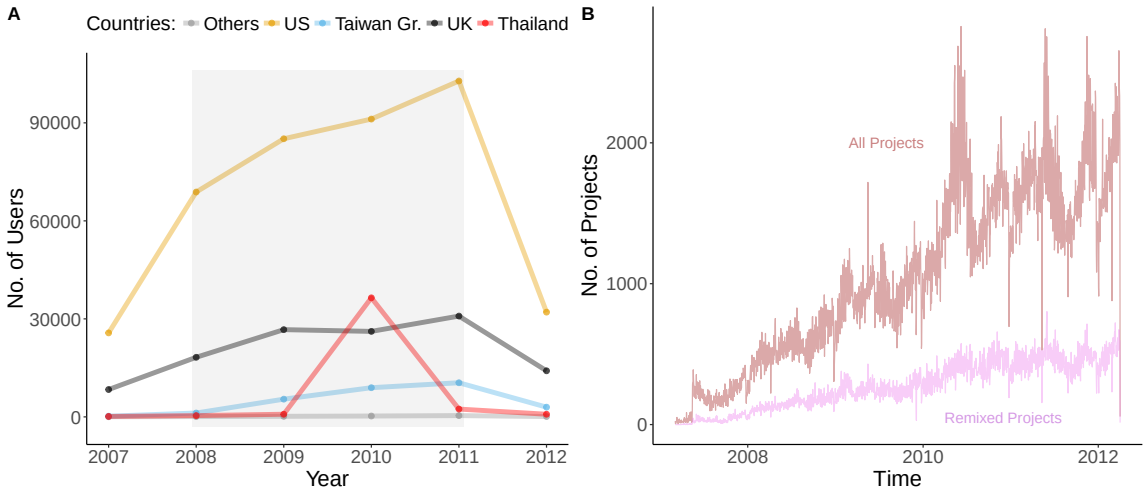
A schematic representation of interactions on the platform as discussed above is shown in Figure 14. Technical details about data quality (missing data, possibly spurious data) are documented in [5]. Wherever required for this study, we discuss the data quality during our analysis.

3 Descriptions of Aggregate Behaviour

This section is composed of descriptive findings of Scratchers' aggregate behavioural patterns during the first five years. First, we see the growth of users and projects on the platform. Second, we provide a manual classification for producers of projects based on their intensity to create remixed projects. Finally, we show that projects can be separated into groups based on their joint consumption patterns and that most Scratchers tend to consume projects from specific groups only.

3.1 Users, Projects

Figure 3: Users Joining & Projects Creation



Panel A shows the joining of 1,056,950 users during each year from March 2007 to March 2012. The evolution is grouped into major clusters of countries. Panel B shows the number of projects created daily upto March 2012. A total of 1,928,699 projects were created during the period.

1,056,950 users joined the Scratch community in the first 5 years (Fig.3A) and 1,928,699 projects were created during this period (Fig.3B). The clusters in Figure 3A are obtained using K-means clustering with five clusters; Taiwan Group is a set of nine countries. The most distinguing trends are born by US and UK, and

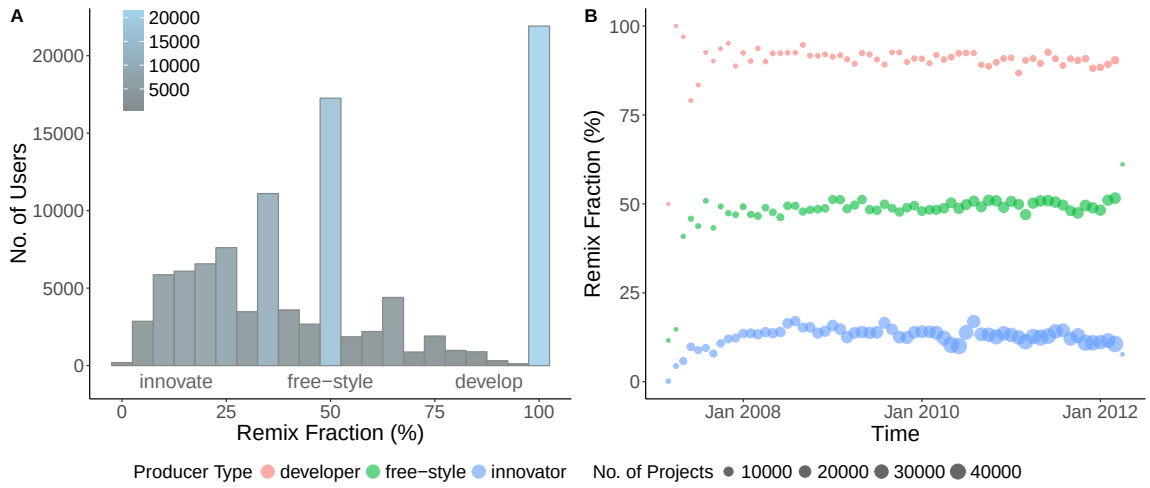
there was a spike in the number of users from Thailand during 2010. We mention some statistics to describe active users on the platform. (i) There are 427,110 users with at least one non-anonymous activity. Since anonymous records include only certain forms of consumption of projects (views, downloads, loves), a large fraction of users in the data are pure consumers. (ii) There are 195,649 users who have interacted (at least one kind of recorded activity) in more than one month (months need not be consecutive). This value does not include users who might have interacted more than one month, but their interactions each month is not recorded (i.e., they only viewed, downloaded, or loved projects). (iii) There are 304,793 users (28%) who created at least one project. This is the sub-population that contributed to the 2 million projects during the five years.

3.2 Production Behaviour

Remixing is a key feature of the Scratch community; it allows creation of projects based on existing ones. Figure 4(A) shows the incentive of Scratchers to remix over the entire time duration; it is a distribution of the fraction of remixed projects among all projects created by a Scratcher. We use this distribution to understand if Scratchers lying in different parts of this distribution differ in certain behaviours. We create three definitions, based on the nature to remix projects: (i) developers: producers with remix fraction greater than 75 %. The value of 75 is chosen to increase the number of users in developers category; most of these users are in the top 10 percentile of the distribution. (ii) innovators: producers who mostly create new projects – producers with remix fraction less than 35 %, (iii) free-style producers: producers who do both. These users are the residuals of segmenting producers as developers and innovators. The value of 35 is chosen such that free-style producers, on aggregate, have 50% new projects and 50% remixed projects. Changes around the cut-off values of 75 and 35 does not affect the number of users in the interval very much. Innovators and developers have contributed to about 87% and 9% of new projects (non-remix projects) respectively.

The definitions are based on aggregate projects (and remixes) created in the entire duration. Users join the community over time, and they fall into one of these producer types (excluding non-producers) as defined by us by looking at the data of entire duration. To see if the labeling of producers based on production in the entire duration also holds in shorter intervals, we calculate the average remix fraction within each type in monthly windows as shown in Figure 4(B). It shows the average remix fraction of each producer type over time. We see that the group of producers categorized as developer type (based on aggregate activity) are of type developer in almost every month: each month, the group produces projects that are mostly remixes. Developers do not tend to behave as free-style producer or innovator in any month, except during the very early period. This suggests that the nature of producers is not volatile, and can be interpreted as a time-invariant behaviour. The distribution of the time spent on the platform by each of these types is almost same, with an average of about 22 months and a standard deviation of 14 months.

Figure 4: Types of Producers



(A) The distribution of remix fraction, i.e., percentage of remixed projects out of all projects created. There are 304,793 users who created at least one project. 202,018 users have zero remix fraction, and are not shown is the plot. The distribution is segmented into three types of producers: innovators, free-style producers, and developers. (B) The average remix fraction during each month for the three types of producers.

The volatility of free-style producers and developers in the early period can be explained by the fact that these producers need existing projects to remix. In the initial periods, since the online community was launched in 2007, there were less projects on board for these producer types to act by their nature. These types show a sharp deviation in favour of their nature, which is due to the availability of more projects in the platform due to passage of time.

Figure 4(A) considers only non-zero remix fractions; users with exactly zero remix fraction are not shown. It forms a large fraction of all producers, however, we are not sure if such producers really did not remix at all. This is because we found some of these users to have produced large numbers of projects in comparision to others (outliers); such a situation might arise by copying projects [23]. Copying projects, although legal, is however unethical; copying is a situation in which a Scratcher modifies a non-substantial part of a project and then posts

it as a new project and not as a remixed project (thereby referencing the original creator). Although copying can arise in other sections of the distribution in Figure 4(A) as well, we did not find evidence of outlier cases for free-style producers and developers. In later analysis and discussions, we therefore study only these producer types: free-style producers, developers.

3.3 Consumption Behaviour

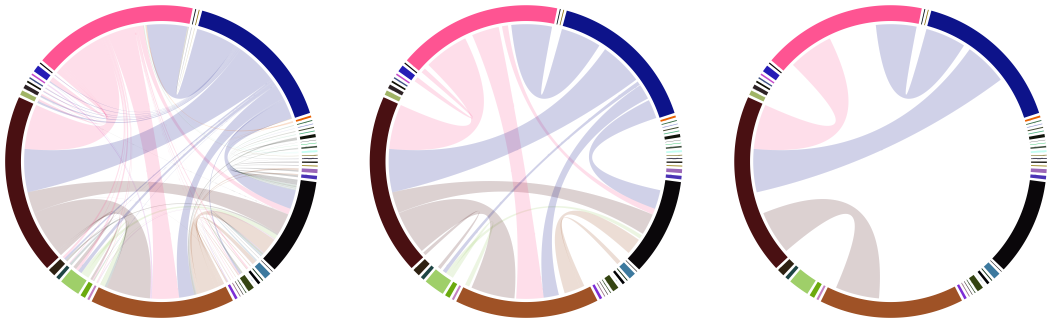
Here we investigate if Scratchers, as consumers of projects, have preference to consume certain kind of projects. First we look for major consumption groups and next we investigate Scratchers' consumption specificity for such groups. Of the various forms in which users consume projects, only favorites and comments are non-anonymous records, i.e, in the available data, we can know the user who favorited or commented on a project. On the platform, favorites and comments are private and public information respectively.

We consider a bipartite network of favoriting behaviour in which the nodes are users and projects, and edges are directed from users to projects which they have favorited. This is an aggregate network considering all favorites interactions in the 5 years. To see which projects are favorited together we obtain a bipartite projection on all projects; in the resulting network, an edge between two nodes (projects) has a weight equal to the number of Scratchers who favorited both the nodes. In the projected network, $\mathcal{P}_{favorites}$, there are 326,975 nodes and 162,611,378 edges with varying weights (ranging from 1 to 442). In the subset of $\mathcal{P}_{favorites}$ with edge weights more than 2 (for simplicity), we found 145 communities in the network by implementing the Louvain algorithm [24]. (We performed communities detection using other algorithms as well, for example, we found 171 communities using fast greedy algorithm [25]. The main results that we discuss below is independent of the choice of algorithm.) 5 among the above 145 communities are of large sizes than others and the inter-community edge densities are low, as shown in Figure 5. We perform a similar community detection on the bipartite network of commenting behaviour. In its projected network, $\mathcal{P}_{comments}$, there are 878,811 nodes and 1,097,722,712 edges, with edge weights ranging from 1 to 323. We found 4 large sized communities, using edge weights greater than 3.

We checked the tags of projects in each community to see if the projects across communities differ by particular topics. We found all communities have similar set of tags – game, simulation, animation, art, music, mario etc. – which are indeed very common tags on the Scratch platform. So the joint consumption of projects does not seem to be segregated by themes (as inferred by tags). We might conjecture that the communities are formed by a Scratchers' location in the friend-

Figure 5: Consumption Communities

(a) all edges (b) $d > 0.001$ (c) $d > 0.01$



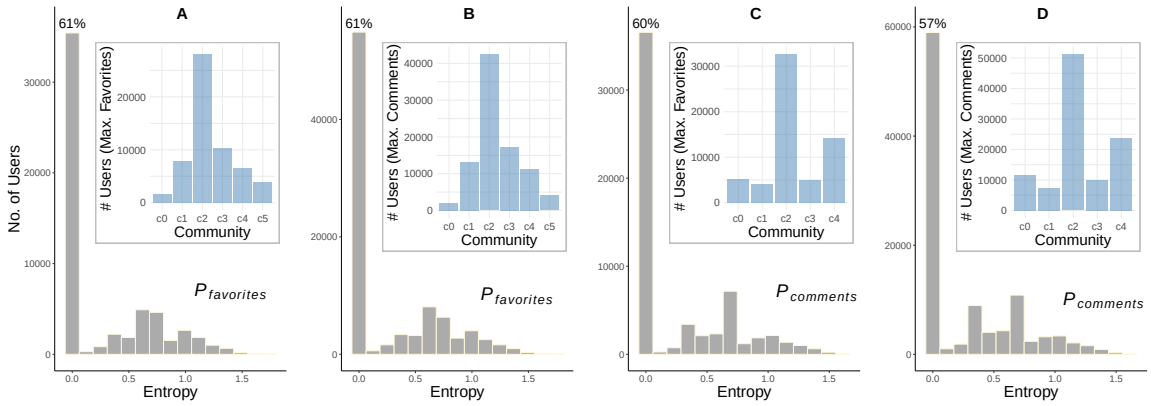
The communities in $\mathcal{P}_{favorites}$ network, obtained by projecting users $[u]$ on projects $[p]$ nodes in the bipartite network where an edge $u \to p$ represents u favorited p. Grids along the circumferences represent communities, and grid size is proportional to community size. Projects within each community are favorited together with high density. Projects of different communities are also favorited together but have low densities d, as shown by the edges between communities. Plots (a), (b), (c) show inter-community densities of minimum values of 0 (i.e., all links), 0.001, and 0.01 respectively.

ship network upon joining the platform, and Scratchers in different segments of the network consume projects of similar themes. To investigate this further, we next examine whether each Scratcher tends (intentionally or unintentionally) to consume projects only from specific communities found above.

We consider Scratchers who consumed (favorites, comments) projects from at least one of the 5 big communities, labelled $c_1, ..., c_5$, found in $\mathcal{P}_{favorites}$. For each of these Scratchers, consider the distribution of consumption across $c_0, c_1, ..., c_5$ where c_0 is the residual community of all projects not included in $c_1, ..., c_5$. We measure a Scratcher's consumption polarization by an entropy-alike measure

$$\mathcal{H} \;=\; -p_0 \log(f(p_0)) - \sum_{i=1}^{n} p_i \log(p_i),\; n = 5; \quad f(p_0) = \begin{cases} 0.5 & \text{if } p_0 = 1, \\ p_0 & \text{if } p_0 \neq 1 \end{cases}$$

Figure 6: Polarized Consumption



Main plots show distribution of entropy $\mathcal{H}$ for consumption of types favoriting and commenting. In each case, there is a high fraction of $\mathcal{H} = 0$, meaning most users consumed projects exactly from a particular community. Insets show the distribution of users' maximal consumption group during 2007-12; for example, the inset in (A) shows that more than 20,000 users favorited projects from the c_2 community in $\mathcal{P}_{favorites}$ the maximum time. (A), (B) show distributions for $\mathcal{P}_{favorites}$ and (C), (D) show that for $\mathcal{P}_{comments}$ network.

where p_i, $i = 0, ..., 5$, is the fraction of consumption from community c_i during the entire duration of five years. It is easy to verify that, with at least one positive p_i, the value of $\mathcal{H}$ is 0 if and only if exactly one value of $p_i, i = 1, ..., 5$ is 1. So if $\mathcal{H} = 0$ for a Scratcher, he has consumed projects from exactly one of the 5 big communities. Figures 6A and 6B show the distribution of $\mathcal{H}$ values of all Scratchers for consumption of types favoriting and commenting respectively. About 60% of Scratchers favorite projects from only one community (Fig.6A), and same is the case for commenting behaviour (Fig.6A). We repeat this analysis considering the 4 big communities ($n = 4$) found in $\mathcal{P}_{comments}$. As shown in Figures 6C and 6D, we find evidence of polarization similar to the case for $\mathcal{P}_{favorites}$. This confirms our earlier conjecture that Scratchers consume projects mostly from specific communities only (the communities are however not different from each other according to the themes of projects within each community).

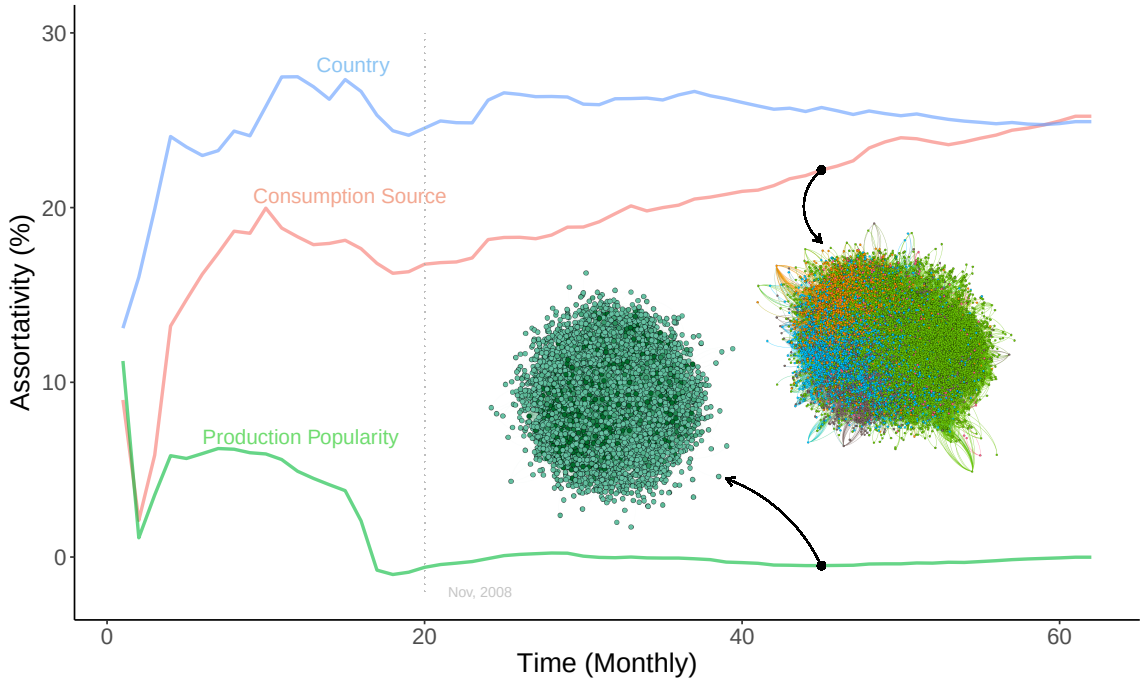
3.4 Assortativity in Behaviour

A friendship network is formed by Scratchers following each other on the platform. This can lead to observations of behavioural similarities among Scratchers and their neighbours. We look at how two attributes – production popularity and consumption preference – are clustered in the network.

We mention the measures used for production popularity and consumption preference. *(1)* There are various observable measures that convey information about popularity of a project. These include love-it, download, and comment. These measures are very correlated because these are determined, often at the same time, by a consumer after he views a project. Favorites count is not observable as a consumption statistic on project page, multiple comments can be made on a project by a single consumer, and downloads count has data issues (the count is supposed to be one per user, but multiple count was found for some users). So we choose love-it as our measure of pouplarity; one consumer can love a project once only. Although there is no platform-specific measure for a project's quality, the love-its received on a project supposedly captures the quality of the project, as assessed by consumers who viewed the project. *(2)* For consumption preference, we study the source of consumption. We use the five big communities in $\mathcal{P}_{favorites}$ as the sources. At a given time, a Scratcher's consumption preference is determined by the source from which he has consumed (favorites) the most.

We measure clustering in behaviour using the assortative mixing coefficients [26], considering numeric and categorical values for production and consumption behaviours respectively. The evolution of the assortativities are shown in Figure 7. The coefficients are significant at 1% level, tested using null model obtained by random shuffling of all edges in the friendship network. (As a comparative reference, the evolution of clustering based on similarity of Scratchers' countries is also shown.) There is no trend for clustering according to production popularity and there is high clustering based on consumption preference. Figure 7 also shows subsets of the network in December 2010. Edges of the subgraph for production popularity are not shown (for clarity of visualization); snapshots are plotted using ForceAtlas2 algorithm [27], so nodes in closer vicinity represent closer neighbours. Scratchers having same consumption sources are clustered

Figure 7: Assortative Mixing in Friendship Network



The evolution of assortativity coefficients for production popularity and consumption source. For a particular month, assortativity is calculated using all edges in the network upto that month. Assortativity for country attribute is shown as a reference. The evolution pattern seems to be stable after November 2008. Insets show subsets of the network at December 2010 to visualize the assortativity values.

and the pattern is dense. Scratchers with similar production popularity are however not clustered – since the measure has many values, we rescaled the colors to have more weight on high values – users with high values of popularity are not clustered and are distributed throughout.

Observations of behavioural clustering in network (Figure 7) can arise due to several mechanisms, including peer influence. To identify the presence of peer influence, other mechanisms that induce clustering in behaviour need to be controlled [28, 29]. We investigate this in the following section.

4 Peer Influence Analysis

Here we investigate the effect of Scratchers' friendship network on their production and consumption of projects. First, we describe our methodology, and next we present the results.

4.1 Methods

Potential Issues: To set the terminology, the focal node or focal actor in a network is called the ego and ego's immediate neighbours are called alters or peers. We want to infer if peers' behaviour influences ego's behaviour. Individuals in a dynamic social network may interact due to many reasons, leading to a simultaneous evolution of network and behaviour of individuals. Some of these reasons established in the literature include [30, 31, 32, 33, 13] exogenous network formation due to homophily and selection, endogenous network formation due to reciprocity and transitivity, peer influence on behaviour, own influence on behaviour, and contexts that lead to certain network-behaviour dynamics. Therefore estimating peer influence using a cross-sectional observation is prone to effects coming from other unwanted reasons, some of which can be of confounding nature. In estimating peers' influence on behaviour, confounding factors are those factors that affect both the ego's behaviour and the peers' behaviour under study. All reasons mentioned above are potentially confounding. An unbiased estimation of peer influence therefore requires control over such confounding factors. Dynamic observations facilitate the separation of co-evolution issues - for example, homophily and influence.

Scenario & Assumptions: A Scratcher's primary activities are producing and consuming projects and galleries. The Scratcher (ego) has the option to follow the creators (peers) of projects – which he likes during random browsing, which he likes in galleries, or interesting projects that appear on the front page of the website. Through activity feeds, the ego knows about consumption (love-its, favorites) and production (projects sharing, remixing) activities of his peers. So we can expect that the ego's activities might be influenced by his peers, in addition to his own tendencies to produce and consume projects. At a given time, the ego can also browse through the projects of his peers to see the production and

consumption statistics. These statistics, like comments count on a project, are the aggregate comments the project has received till this moment. So at a given time, we can assume that the Scratcher has knowledge about his peers and aggregate statistics of their activities (total projects, total loves received, etc.).

In measuring peer influence at a given time, we assume that the Scratcher is influenced only by the aggregate activities of his peers upto this time and not by the history of such activities. It is very unlikely that a Scratcher remembers the exact history of previous activities of his peers. For example, consider a Scratcher with just one peer, and we are interested to investige the peer influence of projects count: we assume that the total projects produced by the peer upto now influences the Scratcher on how many projects he produces next, and he is not particularly influenced by the exact number of projects his peer produced during the last week (or any particular historical period in general). This behavior is termed mathematically as Markovian. Markov nature of decision making is a very plausible assumption in the scenario of Scratch community. This property has been widely adopted in the social networks literature - for example, in stochastic actor oriented models [11], future decision of network or behaviour change made by an actor is conditioned on the network and behaviour in the present state. Essentially, under Markov assumption all variables (network, behaviour) of interest are represented as state variables at the time peer influence is evaluated.

Quasi-experiment: We define peer influence of a behaviour b_{peers} at a time t on ego's behaviour b_{ego} at time $t + j$ $(j = 1, 2, ..)$ as the exogenous influence of peers' state of behaviour (known to ego) at t, b^t_{peers}, on ego's state of behaviour at $t + j$, b^{t+j}_{ego}. State variables at t summarize behaviours upto time t and form the basis of ego's decisions at t (Markov nature). To measure peer influence, treatment status is assigned using a binary variable Tr^t which is based on a threshold value of b^t_{peers}. All Scratchers (entire population) at t are distributed into two groups: treated and control; Scratchers in the treated group ($Tr^t = 1$) have high values of b^t_{peers}, and those in the control group ($Tr^t = 0$) have low values of b^t_{peers} and serve as the counterfactual. At this point, treatment is likely to be correlated with several confounding variables, and so treatment effect estimates would be biased. So we obtain a subset of the population at t, by matching exactly on confounding

variables, such that treatment can be justified to be randomly assigned across treated and control groups in the subset. In this sub-sample, having controlled for possible confounding effects, we capture the effect of treatment on change in behaviour of treated group ($\Delta b_{ego}^{t \to t+j} = b_{ego}^{t+j} - b_{ego}^{t} \mid Tr^t = 1$) and compare it with the counterfactual effect ($\Delta b_{ego}^{t \to t+1} \mid Tr^t = 0$). Peer influence at t is thus measured as the difference of the future changes in behaviour b_{ego} across treated and control groups. This forms the basis for peer influence estimation under a quasi-experimental setting. Below we present the empirical implementation and discuss the validity of our method.

Implementation: We employ an ego-centric regression framework to assess, at time t, the impact of being treated on ego's future change in behaviour.

$$\mathbf{\Delta b}_i^{t \to t+j} = \alpha^j + \beta_{peer}^j \mathbf{Tr}_i^t + \underbrace{\beta_1^j \mathbf{N}_i^t + \beta_2^j \mathbf{X}_i^t}_{\substack{\text{confounders are} \\ \text{balanced across } Tr(1,0)}} + \epsilon_i^{t \to t+j},\ j = \{1, 2, 3, ...\} \quad (1)$$

$\Delta b_i^{t \to t+j}$ is the change in behaviour b of ego i from time t to $t+j$. All explanatory variables represent behavioural state at t, measured as an aggregate operation on observed behaviour upto t. For example, to represent the comments behaviour of ego at t, we use the total comments made by the ego upto t as the state variable at t. Tr_i^t is the treatment variable – the variable of interest that represents peers' behaviour at t. It is a binary variable – values 1 and 0 are assigned to egos in the treated and control groups respectively. Under absence of selection bias, β_{peer}^j represents the average treatment effect on change in future behaviour. Treatment status Tr_i^t for ego i can change over time (i.e., a Scratcher who is in the treated group today can be in the control group at another time) because the ego is assigned to either treated or control group based on peers' behavioural state at t. Hence estimates β_{peer}^j are conditional on time t. N_i^t represents ego's network variables at time t. In general, it can incorporate information of the entire network of ego upto neighbours at any distance. However, for practical purpose it is sufficient to include characteristics of ego's local network (immediate neighbours) only – structural properties of ego's network (e.g., out-degree, in-degree, reciprocity), various behaviours of peers (excluding the behaviour represented by the treatment variable), and structural properties of peers' local network. X_i^t

represents various characteristics of the ego at time t. It includes the dependent behaviour b under study as well to capture auto-correlation of behaviour, or in other words, ego's own tendency. In this study we use monthly windows – t is the time at a month's end, $t+1$ is the time at the end of next month, $t+2$ is the time at the end of 2 subsequent months from t, and so on. In selecting covariates X_i^t and N_i^t, we employ these criteria: (i) the variable should be, conceptually, a potential confounder, i.e., it can affect both Tr_i^t and $\Delta b_i^{t \to t+j}$, (ii) peers' variables should be such that they can be assumed to be known to the ego; for e.g., projects creation by peers appear in ego's activity feed as an information, (iii) not include new variables that are simultaneously determined and signify same behavioural information as included variables, and (iv) variables with high multi-colinearity during estimation of the regression model are excluded.

Exact matching is used as a preprocessing step prior to estimating (1) in order to achieve balance in confounding variables across treated and control groups. Regression analysis following the matching step leads to statistically consistent estimates [34]. We implement one-to-many exact matching, in which each treated unit is matched to multiple units in the control group having exactly the same values of the matched variables [35, 34]. Each matched control unit has weight proportional to the number of treatment units to which it is matched, and the sum of the control weights is equal to the number of uniquely matched control units. Unmatched units have weights equal to 0, and matched treated units have weight 1. The regression analysis that follows matching uses weights corresponding to each unit produced during matching stage [34]. Exact matching costs data, so we exploit high correlations among variables to obtain balanced samples by matching only on subsets of (important) confounders. Eventually, for regression analysis, we use samples which produces the best balance (reduces bias in regression estimates), and also retains a good sample size (reduces variance of regression estimates). Variables which remain unbalanced are controlled at the regression stage. Balance in model variables in (1) is assessed by difference in weighted means of variables across treated and control groups. Balance is assessed on all model variables, including those that are excluded from matching analysis. For a given matched sample, with a good balance of the post-matched variables across treated and control groups, regression estimates are not supposed to change dramatically across different models.

We summarize the practical steps involved in estimating peer influence using the method presented above. *(i)* determine model variables, i.e., all potential confounders *(ii)* dichotomize treatment, if needed *(iii)* determine selection into treatment, i.e., statistically relevant confounders *(iv)* match exactly on (subset of) confounders to achieve balance *(v)* estimate model using OLS.

Internal Validity: We want to obtain unbiased estimates β^{j}_{peer} of the treatment effect in (1). Since we do not expect reverse-causality issues, selection bias is the most important source of bias. This arises due to factors that affect both the treatment and future change in ego's behaviour, and are not affected by the treatment itself or by anticipation of treatment. Same intensity of selection bias across treated and control groups can justify a random assignment of treatment, and minimize alternative mechanisms. *(a)* Exact matching as a preprocessing step and including controls in the regression stage help to minimize selection bias due to observable factors. We control for exogenous network formation processes, homophily and selection, by accounting for balance in ego's characteristics and peers' characteristics [13]. Endogenous network formation processes like general tendency to follow Scratchers (out-degree) and tendency to follow one's followers (reciprocity) can be confounding, so we control for such factors as well. Change in future behaviour can also depend on the level of behaviour at t, and this is a major confounder because Scratchers in the treated group are more likely to have higher behavioural levels because of correlation between peers' behaviour and ego's behaviour. We therefore always include this factor in all matching analysis; the sub-samples have exact levels of behaviour across treated and control groups before the onset of change in behaviour. *(b)* Comparision of future change in ego's behaviour with a counterfactual group takes care of selection into treatment due to unobservables factors, as long as such factors are time-invariant. *(c)* The treatment variable is dichotomized [34] before matching analysis according to certain thresholds that suggest high or low levels of behaviour. We need to analyze peer influence for various thresholds to ensure that peer influence estimates are not extremely sensitive to such choices of thresholds.

4.2 Results

In this section we analyze peer influence for two different behaviours. First, we analyze an attribute that represents behaviour of Scratchers' production. For production behaviour, we study the popularity effect – whether the popularity of projects created by a Scratcher increases (or decreases) if his peers' projects are popular. Second, we analyze an attribute describing Scratchers' consumtion behaviour. For consumption behaviour, we study the preference effect – if peers of a Scratcher consume (by favoriting) projects from a certain source, does the Scratcher tend to consume projects from the same source in future? For estimating peer influence, we follow the steps mentioned in previous section (Methods). We found presence of statistically significant peer influence only for production behaviour. In the remaining part of this section, we provide empirical details of peer influence for production behaviour and an outline of the empirics for consumption behaviour.

Production Popularity A Scratcher accumulates love-its on a project he created when another Scratcher (consumer), either his follower or a random user, who views the project finds it interesting (most likely due to project quality) and clicks the love-it button. If the project receives a lot of attention (as inferred by love-its, comments, downloads, etc.), it can be selected to appear on the front page. This selection can be system-based or by admins. The project can also appear in some galleries. These would most likely increase viewership for the project, and the project is subject to more love-its. Figure 15 shows a schematic diagram of the process of accumulation of love-its. All factors that affect the quantity and quality of projects created by a Scratcher, and the total views on all his projects are predictive of his popularity (measured as total love-its). In reference to model (1), these factors are of types X_i^t (ego's attributes) and N_i^t (ego's local network structure, peers' observable characteristics, peers' local network structure), and are selected acording to the criteria mentioned in previous section (Methods). The measures of covariates X_i^t and N_i^t represent the respective behavioural states at the time t of peer influence evaluation (see Table 2).

The treatment variable of interest is popularity of peers' projects at t (Tr_i^t); since it is not a binary measure, we dichotomize as

$$Tr_i^t = \begin{cases} 1 & PQ_i^t \in (c_{min}, c_{max}] \\ 0 & PQ_i^t \in [0, c_{min}), \end{cases} \tag{2}$$

where

$$PQ_i^t = \sum_{j \in \{\text{peers of ego } i\}} (\text{total love-its on all projects upto } t)_j$$

is the measure for peers' popularity for ego i at time t, and c_{min} and c_{max} are self-chosen values representing minimum and maximum threshold values respectively. (We shall see later how the chosen thresholds affect the results.) Our dependent variables of interest are future changes in production popularity of ego i:

$$\Delta b_i^{t \to t+j} = b_i^{t+j} - b_i^t, \; j = 1, 2, ..,$$

where b_i^s represents the total love-its accumulated by all projects of ego i upto time s. Peer influence estimation is conditional on time t; to have sufficient observations, we begin by analyzing in the stable period of the data: the month of December, 2010. (t represents the end of Dec, 2010 and $t+1$ is the end of Jan, 2011.) We use the median value of peers' popularity at t as c_{min} and the maximum value of PQ^t as c_{max}.

In Table 3 (columns 1,2), we see that almost all variables (included in model as potential confounders) are valid model variables. To learn which of the potential confounders are important, i.e., affect treatment variable and hence can lead to selection bias, we perform logistic regressions of treatment Tr_i^t on covariates X_i^t and N_i^t. As shown in columns 3 and 4 of Table 3, we perform two such regressions: first with only X_i^t variables and next with all variables. Having determined the significant confounders from this analysis, we perform exact matching on (all or subset of) such confounders across treated ($Tr_i^t = 1$) and control ($Tr_i^t = 0$) groups with an immediate goal to have balance of all variables – includes all potential confounders, irrespective of their statistical significance in logistic regression – across both groups. Balance is determined by the difference in weighted average values of the variables in each group. We obtain two reduced datasets:

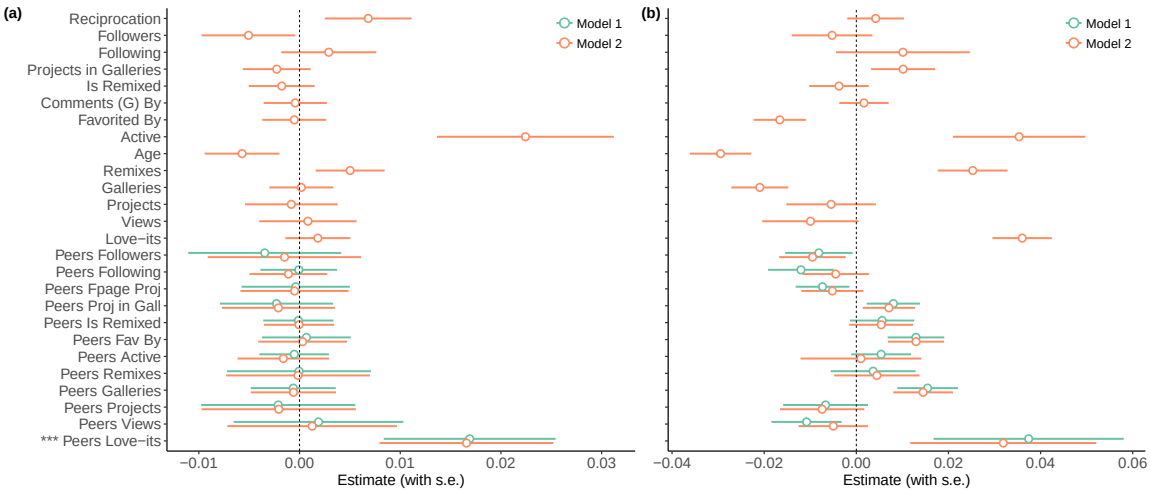
(a) sample obtained by exact matching only on all X_i^t variables that are significant (Table 3, col. 3). This is done to (i) compare our method with Aral [20], which matches on all individual (ego) characteristics, and (ii) show that our results are robust to matching strategies that produce good balance,

(b) sample obtained by exact matching on a subset of all (X_i^t and N_i^t) variables that are significant (Table 3, col. 4). Matching exactly, especially with N_i^t variables was found to be very costly, and so only few variables were used. Among all combinations of variables we investigated, we present the one with best balance. This sample is less biased than the one in (a), and also has more obervations because matching is performed on less covariates.

We show balance for all variables (difference in averages of variables by treatment groups) in Table 4. Although we do not use, we show balance produced by matching via propensity score method, as used by [20]; this method produced almost no improvement (as compared with the original imbalance in the full sample) in covariates balance. Having balanced samples, we can assume we are in a scenario where treatment (high popularity of peers) has been randomly assigned to each Scratcher (ego). The quantitatively small imbalances that still remain for some covariates are controlled during the regression stage.

Figure 8 shows the estimated coefficients for (1) with $j = 1$, i.e., the effect of peers' production popularity on Scratchers' production popularity the next period. Estimates are shown separately for two cases, corresponding to the two reduced datasets. For each matched sample, we see that the estimated model coefficients are stable across various specifications and the peer influence coefficient $\beta_{peer}^{j=1}$ is positive. In both cases, β_{peer}^{1} is significant at 1% level. Details of regressions corresponding to (a) and (b) in Figure 8 are available in Table 5. The identification of β_{peer}^{1} relies on random assignment of observations to treated and control groups. Achieving a good balance of covariates across both groups and including covariates as controls in regressions minimizes selection bias to a large extent. However, another source of non-random treatment assignment lies in the definition of Tr_i^t variable which depends on the chosen values of c_{min} and c_{max}. The results in Figure 8 use one pair of values; so we need to check whether peer influence estimate remains significant and how it varies when threshold values

Figure 8: Peer Influence on Next Month

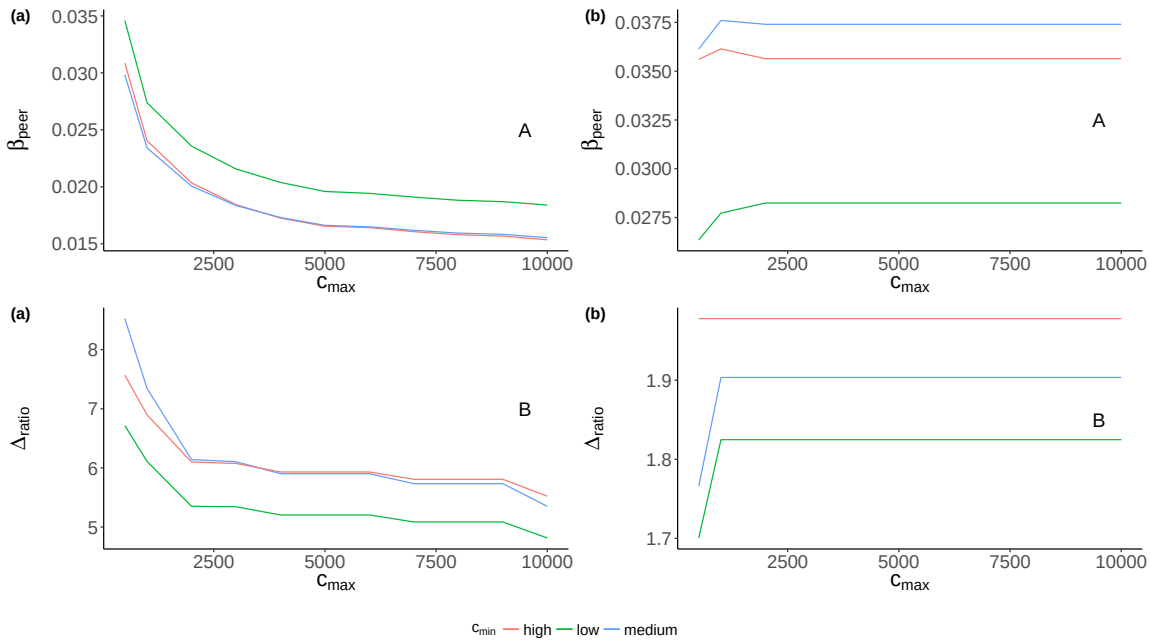


Estimates for regression (1) using samples obtained by matching on (a) X variables and (b) both X and N variables (t = Dec 2010, $j = 1$). In (a) and (b), Model 1 includes only peers' attributes (N), and not egos's attributes (X), as controls. The effect of peers' production popularity (Peers Love-its) on Scratchers' change in production popularity next month, β_{peer}, is significant at 1% level. See Table 3 for variables description, Table 4 for details of matched samples, and Table 5 for details of estimates.

c_{min} and c_{max} change. Figure 9 shows this robustness analysis. Since β^1_{peer} is the primary coefficient of our interest (peer influence), we plot these estimates for changing threshold values, as shown in panels labelled A. All estimates of β^1_{peer} are postive and significant at 1% level. For each value of c_{min}, peer influence estimate tends to decrease with increase in c_{max}. Panels labelled B show the ratio of weighted means of outcome $\Delta b^{t \to t+1}$ variables in treated and control groups, a measure of relative comparision of outcomes without any post-matching adjustment for confounders.

Above, we provided details of the peer effect of production popularity in the immediate period ($j = 1$) using the network existing at the end of December 2010 (t). Popularity of peers' projects at time t might influence the popularity of a Scrather's (ego) projects in subsequent periods as well; in model (1) this

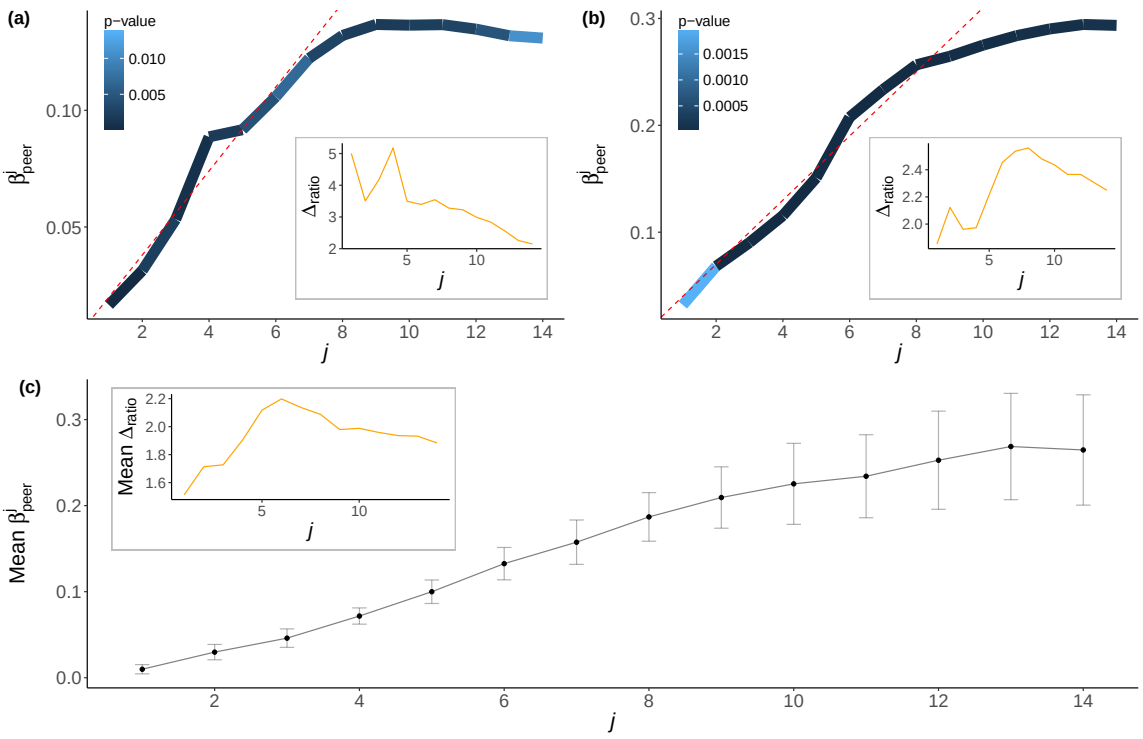
Figure 9: Peer Influence on Next Month: Robustness Check



Effect of varying threshold values of c_{min} and c_{max} on estimates evaluated on samples obtained by matching on (a) X variables, and (b) both X and N variables (t = Dec 2010, $j = 1$). Panels A: Estimate of β_{peer}, controlling for both X and N variables in the regression. Panels B: Ratio of average of future changes in production popularity of treated and control groups.

corresponds to values of j as $2, 3$, and so on. To see if there is persistence of peer influence in subsequent periods, we estimate β^j_{peer} for various j by changing the dependent variables in (1). Since we are looking at the effect of peers' popularity at time t on period $t + j$, the treatment assignment Tr^t_i, and hence the reduced datasets obtained by exact matching, remains the same as in the analysis for Figures 8 and 9. The results for persistence of peer effect are shown in Figures 10(a) and 10(b). The effect of peers' popularity at t on Scratchers' (ego) production popularity at a future period $t + j$ increases with subsequent periods. The rate of increase tends to steep up during the middle term and tends to flatten out in the long term, thereby creating a S-shape for the structure of persistence curve. This

Figure 10: Persistence of Peer Influence



(a) Estimates of β^j_{peer}, the effect of peers' production popularity at t = Dec 2010 on production popularity at future periods $t + j$, for varying j. Inset shows the ratio of future changes in production popularity of treated and control groups. Sample in Table 4*a* obtained by matching egos' characteristics X^t is used for evaluation. *(b)* Same estimators as in (a), using sample in Table 4*b* obtained by matching individual and peers' characteristics. (*c*) A general persitence curve, showing the average β^j_{peer} during July-Dec 2010, i.e., for each j-periods ahead influence, the plot shows its average value calculated for each month during July-Dec 2010. Error bars are scaled standard deviations of β^j_{peer}. Inset shows the ratio of mean future changes in production popularity of treated and control groups. Sample in Table 4*b* is used.

shape is more prominent in Figure 10(b), compared to 10(a), where the balance in underlying matched sample is better (Table 4).

Persistence curves at two different times are ideally not comparable because treatment assignment on Scratchers (ego) can vary from one period to another. So performing peer influence analysis at a time $\tilde{t}$ different from the one we have used in above analysis (t = Dec, 2010) requires obtaining a balanced, preferably the least biased, sample at $\tilde{t}$ (by matching exactly on confounders that are significant at $\tilde{t}$, which may differ from those at t). However, assuming that the users' behaviour to be stable over the last six months of 2010, we can assume that the selection into treatment in each of these months (t) follows a similar pattern as we saw above. So we create reduced samples by exactly matching on the same set of variables as in column (b) of Table 4. We expect that the imbalances in other months would be more than that in Table 4 (which corresponds to month of December); however controls in the regressions help to reduce bias. Now we look at the persistence curves $\beta^{j}_{peer}(j = 1, ..., 14)$ for each of the last six months (t = July 2010, ..., Dec 2010); the average of these curves is shown in Figure 10(c). For each j, we plot the average of β^{j}_{peer} estimates obtained in six different models (1) corresponding to six different values of t.

We saw earlier that the assortative mixing coefficient for projects popularity is near zero, which means that there is no observed clustering based on popularity of the projects, and this is consistent over time. However, we do find positive short term and long term peer influence, which means that a positive measure of clustering can be expected at a future time when the effect of influence has taken place. Of several probable mechanisms that might explain this, we provide a qualitative example to illustrate the main idea. Consider a situation, at time t, where an ego has four outgoing friendships, two of which have higher production popularity than him and the other two have lower popularity than him. This suggests that ego's local network is not assortative based on this behaviour. During time t to $t + 1$, the ego receives a small increase in its projects' popularity due to influence of his peers. However since his friends also undergo similar change (by influence from their peers), the relative values of popularity in ego's local network at $t + 1$ remains the same as in time t – two friends have better popularity and other two have lower popularity, suggesting a zero value for assortative mixing once again.

Table 1: Heterogenous Influence / Susceptibility

	Immediate (j=1)	Medium Term (j=6)	Long Term (j=12)
		(a) Matched on X	
2nd order effect			
σ (Peers' Popularity)	(0.019 *** , 0 , 0)	(0.085 ** , -0.02 , 0.04)	(0.111 ** , -0.07 , 0.1)
Activity Frequency			
Active	(0.006 , 0.01 ** , 0.03 ***)	(0.034 , 0.16 *** , 0.24 ***)	(0.05 , 0.27 *** , 0.29 ***)
Age	(0.018 *** , 0 ** , 0)	(0.128 *** , -0.02 ** , -0.01)	(0.151 *** , -0.03 *** , -0.01)
Producer Type			
Developer	(0.014 *** , 0 , 0.16 ***)	(0.104 *** , 0.28 , 0.07)	(0.13 *** , 0.38 * , 0.25)
Free-style	(0.017 *** , 0.08 *** , -0.01)	(0.056 , 1.44 *** , 3.51 ***)	(0.082 * , 2.65 *** , 3.59 ***)
		(b) Matched on X, N	
2nd order effect			
σ (Peers' Popularity)	(0.036 *** , 0 * , 0)	(0.207 *** , 0.01 , 0)	(0.303 *** , 0.01 , -0.03)
Activity Frequency			
Active	(0.016 , 0.03 *** , 0.02)	(0.04 , 0.16 *** , 0.24 ***)	(0.038 , 0.24 *** , 0.37 ***)
Age	(0.037 *** , 0 *** , 0)	(0.207 *** , -0.01 *** , 0)	(0.311 *** , -0.02 *** , 0)
Producer Type			
Developer	(0.021 ** , -0.05 *** , 0.23 ***)	(0.196 *** , -0.15 *** , 0.23)	(0.28 *** , -0.25 *** , 0.2)
Free-style	(0.032 *** , 0.05 *** , 0.01)	(0.131 *** , 0.12 *** , 0.92 ***)	(0.198 *** , 0.26 *** , 1.14 ***)

(*i*) In a given row, each tuple (a, b, c) represents, in order, the coefficients of treatment (peers' production popularity), attribute (corresponding to the row name), and the interaction of treatment and attribute. The interaction variable captures heterogeneity of treatment. (*ii*) p–values: significant at 10% level (*), 5% level (**), 1% level (***)

Next we study if particular Scratchers, due to their own nature or local network characteristics, are more susceptible to influence from their peers compared to other identical Scratchers. For this, we use an interaction term (of the desired attribute) with the treatment variable in regression model (1), controling for all confounders as before. The results are shown in Table 1; in each tuple, the first, second, and last elements correspond respectively to the estimated coefficients of treatment variable, desired attribute, and the interaction of treatment and attribute. The first attribute is a network characteristic: variance of peers' production popularity; we see that the interaction term is not significant, i.e., an ego in the treated group with two peers having average production popularities will be influenced in the same way if one of his peers had a high popularity and the other

had a low popularity. So a treated Scratcher is not influenced by specific peers (in general), rather the influence stems from the overall production popularity of his local environment. Not shown here, we tested for several other attributes of peers (total remixes, favorites, projects in front page, etc.) and found no evidence of influence heterogeneity. So, peers' behaviour does not seem to create extra susceptibility, which seems intuitive, because incorporating more influence in comparision to identical others should arise out of individual traits. We see that active Scratchers are influenced more than if not active, and the effect on future periods is increasing; this shows that interacting on the platform for more than a month is required for being influenced by peers. (Not active users can be of two types – those who joined much before but interacted less than one month, and those joined just one month prior to t, i.e., during December 2010.) Further activity frequency does not have a significant effect among treated Scratchers, as seen from the coefficient of age; so it seems that, on average, a prolonged (over one month) interaction either initially or somewhere during the lifetime on the platform is necessary to become more susceptible to the popularity of peers. Remixing is an important characteristic in the learning process in Scratch community. We see that developers are more susceptible to be influenced in the very immediate period, but not in the medium or long term, compared to other producers whose peers have high production popularity but they are either free-style producers or innovators. Most probably, it is by the nature of remixing – if a developer sees popular projects of his peers, he builds on top of it to have new projects in the next period and gain popularity, but he is not influenced by today's production of peers to create projects in the subsequent periods. On the other hand, influence of peers' popularity on free-style producers takes effect in the medium to long term and is very significant. So having traits of innovation, i.e., creating new projects, leads to additional influence from production popularity of peers in the long run.

Consumption Preference The next behaviour that we examine for peer influence relates to consumption. In Figure 7 we saw that Scratchers having the same source of consumption tend to cluster in the friendship network. We investigate to what extent this can be explained by influence from peers, i.e., if peers tend to favorite (consume) projects from certain source, does the ego also tend to increase consumption from the same source? This investigation is relevant because the ego knows about his peers' favoriting patterns via activity feeds (comments made

are not visible to followers via activity feeds), but does not know the 'source' of consumption because the sources have been identified by us by clustering projects over the observed data of the entire duration. Evidence of peer influence in this case would imply that tastes or preferences for consumption of ego are not static and can be affected by peers' preferences.

We consider the group of projects in c_2, the biggest community in the $\mathcal{P}_{favorites}$ network. We examine whether ego increases his consumption of projects from c_2 group if his peers mostly consume projects from c_2 group. In terms of model (1), $\Delta b_i^{t \to t+j}$ is the change in consumption of c_2 projects from t to $t+j$, and Tr_i^t for ego i is 1 only if more than 50% of his peers have consumed projects from c_2 group the maximum time (upto t). Controlling for various confounding factors, we did not find significant coefficient for β_{peer}^j. So we can not conclude that Scratchers are influenced by their peers' consumption interests. The most likely reasons for the observed clustering in Figure 7 therefore seems to arise out of contextual friendship formation among the users, context being the position in the network after initial interactions on the platform, which is followed by consuming projects within large communities locally. Since projects communities contain similar themes of projects (see section 3.3), it is a probable explanation why Scratchers are not influenced by the consumption patterns (favorites) of their peers.

5 Mechanism of Peer Influence

In the previous section we saw that production popularity of peers has a positive effect on the outcome of Scratchers' future production popularity (Figure 10). However this finding does not suggest *how this effect mediates to outcome* – especially, whether any particular changes in activities made by Scratchers in subsequent periods due to popularity of peers' projects affects their future production popularity. The aggregate love-its accumulated by a Scratcher upto a certain time t depends on the projects created by him and the views received on his projects upto that time. The change in production popularity in next periods, i.e., during t and future times $t + j$, can therefore arise due to:

- *(Channel 1)* change in the number of projects created in next periods, which leads to new views and the possibility to have more love-its, and
- *(Channel 2)* change in the number of views in next periods on projects already created by time t, which can lead to more love-its.

Of these two (major) mechanisms of peer influence, the outcome (change in popularity next period) caused via creation of new projects is particularly noteworthy. This mechanism gives better insight into the Markov decision-making nature of Scratchers – upon having popular peers, they are influenced to create more projects (and possibly of better quality) which enhances their popularity in future periods. We denote b_i^t, Tr_i^t, and M_i^t as the production popularity, treatment assignment, and total projects respectively at time t for Scratcher i. Our goal is to disentangle the effect mediated via creation of new projects (channel 1):

$$\underset{\text{peers' popularity}}{Tr_i^t} \xrightarrow{\text{decision}} \underset{\text{new projects}}{\Delta^j M_i} \xrightarrow{\text{views}} \underset{\text{gain in popularity}}{\Delta b_i^{t \to t+j}(Tr_i^t, \Delta^j M_i)}$$

from other mechanisms of treatment effect (channel 2), where

$$\Delta^j M_i \equiv \Delta^j M_i(Tr_i^t) \;\; := \;\; \Delta M_i^{t \to t+j}(Tr_i^t) \;\; = \;\; M_i^{t+j}(Tr_i^t) - M_i^t$$

is the total projects created by Scratcher i during t and $t + j$ (the mediating variable of interest) and is a function of peers' production popularity at t (Tr_i^t).

To do so, we employ model-based causal mediation analysis [36] where the treatment variable Tr_i^t is randomized, conditional on the confounders X_i^t and N_i^t (quasi-experimental setting), and the mediating $\Delta M_i^{t\to t+j}$ and outcome $\Delta b_i^{t\to t+j}$ variables are observed without interventions. Estimates of β^j in regression (1) is the total average treatment effect *ATE(j)* [3], and is the sum of average effects of treatment at t on gain in production popularity from t to $t+j$ via all possible mediating channels. *ATE(j)* is decomposed into:

- *ACME(j)*: the component *of ATE(j)* mediated via creation of new projects (channel 1) is called the *average causal mediation effect* and is defined as:
  $\bar{\delta}^j(Tr^t) \;=\; \mathbb{E}_i\left[\Delta b_i^{t\to t+j}\left(Tr^t, \Delta^j M_i(1)\right) - \Delta b_i^{t\to t+j}\left(Tr^t, \Delta^j M_i(0)\right)\right]$, and
- *ADE(j)*: the component *of ATE(j)* mediated via all other mechanisms (channel 2) is called the *average direct effect* and is defined as:
  $\bar{\xi}^j(Tr^t) \;=\; \mathbb{E}_i\left[\Delta b_i^{t\to t+j}\left(1, \Delta^j M_i(Tr^t)\right) - \Delta b_i^{t\to t+j}\left(0, \Delta^j M_i(Tr^t)\right)\right]$. [4]

To contrast our investigation with the previous results, we fix t at December, 2010 and vary j from 1 to 12. Following [36, 37], we estimate *ACME(j)* and *ADE(j)* in the reduced dataset obtained by matching all variable types (Table 4(b)). Linear models (3) and (4) are used for mediating and outcome variables respectively

$$\Delta M_i^{t\to t+j} \;=\; \gamma_0^j \;+\; \gamma_1^j\, Tr_i^t \;+\; \gamma_2^j\, N_i^t \;+\; \gamma_3^j\, X_i^t \;+\; {\epsilon_M}_i^{t\to t+j} \tag{3}$$

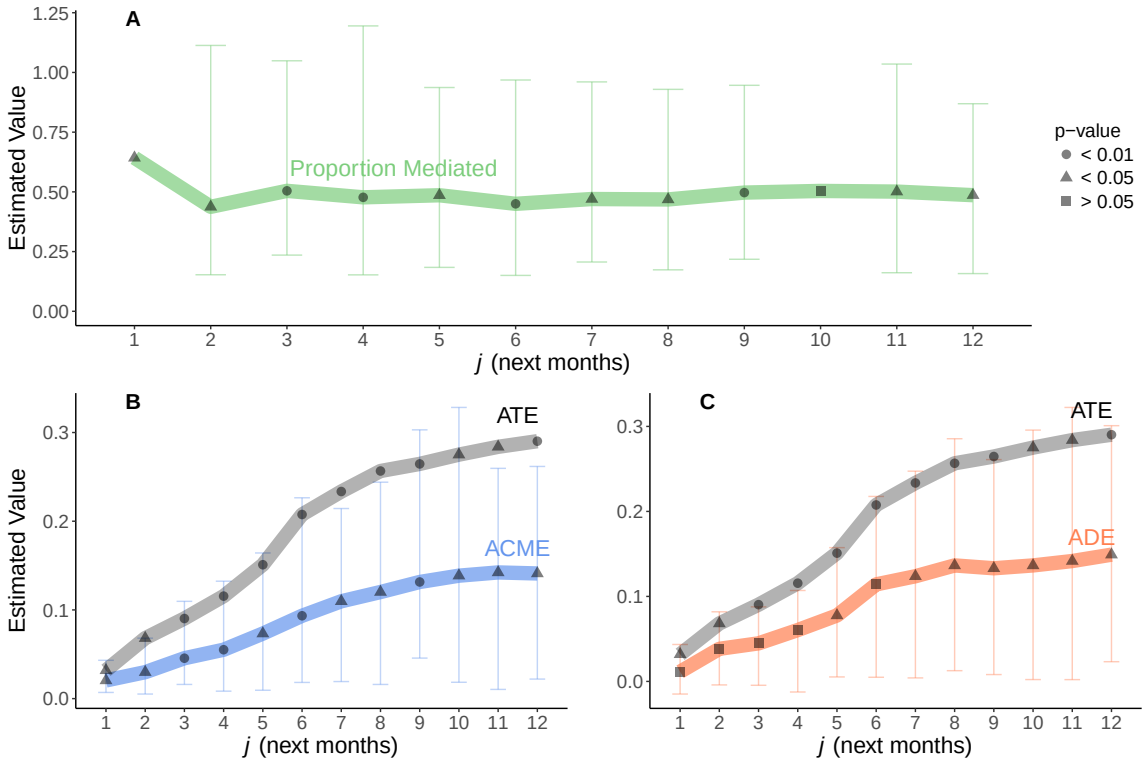
$$\Delta b_i^{t\to t+j} \;=\; \beta_0^j \;+\; \beta_1^j\, Tr_i^t \;+\; \beta_2^j\, \Delta M_i^{t\to t+j} \;+\; \beta_3^j\, N_i^t \;+\; \beta_4^j\, X_i^t \;+\; {\epsilon_b}_i^{t\to t+j} \tag{4}$$

$$\rho^j \;=\; Corr({\epsilon_M}^{t\to t+j}, {\epsilon_b}^{t\to t+j}) \tag{5}$$

where ρ^j is the correlation between the error terms, and X^t, N^t are the same confounding variables as used in model (1), for estimating peer influence on production popularity, to ensure a random assignment of treatment. The estimates and their confidence intervals are obtained using non-parametric bootstrap with percentile method. Figure 11 shows the estimates of *ACME(j)* and *ADE(j)* in pan-

3. β^j in regression (1) estimates the average treatment effect on the treated group ($Tr^t = 1$) and is equal to the average treatment effect (ATE), for the population, when Tr^t is randomly assigned.
4. Note that *ACME(j)* and *ADE(j)* are defined using potential outcomes framework and hence contain counterfactual values. For example, for a Scratcher with $Tr_i^t = 1$, $\Delta M_i(0)$ is not observed because it is the number of projects he would have created from t to $t+j$ if he were assigned $Tr_i^t = 0$. Counterfactual values are estimated from the data during the estimations of *ACME(j)* and *ADE(j)*.

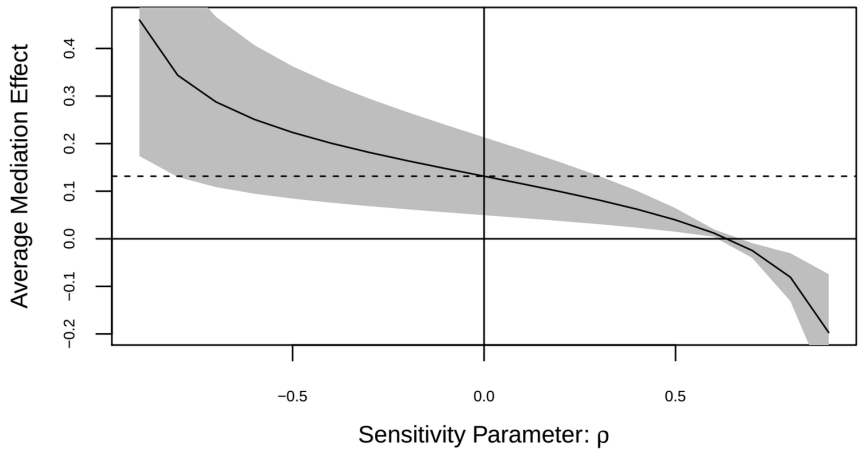
Figure 11: Peer Influence Channel: Creation of Projects



(A) The proportion of peer influence in production popularity, as estimated in Fig.10(b), which is mediated via Scratchers' creation of new projects in future periods j. 95% CIs are shown for each estimate, except at $j = 1, 10$ where the upper boundary is more than 1.25. (B, C) The decomposition of the total average effect of peer influence (ATE) into the primary channel of creation of new projects (ACME), and secondary channel which includes all other pathways (ADE). 95% CIs are shown for ACME and ADE estimates.

els B and C respectively. We see that both estimates are positive and increasing for all j. Both the effects tend to increase at a decreasing rate, and so we observe a similar additive effect for *ATE(j)*. Also, the S-shape for *ATE(j)* seems to arise from *ADE(j)*. The ATE curve in Figure 11 (obtained here using non-parametric bootstrap estimation) is identical to the curve in 10(b). We performed a heterogeneity test [37] with the null hypothesis $\bar{\delta}^j(1) - \bar{\delta}^j(0) = 0$ and concluded that $\bar{\delta}^j(1)$ and $\bar{\delta}^j(0)$ are statistically not different (high p-values $\forall j$); so the *ACME(j)* curve in Fig-

Figure 12: Peer Influence Channel: Robustness Check



The variation of $ACME(j = 9)$, along with 95% CI of the estimate, to confounding effects, as captured by the sensitivity parameter ρ (see (5)). Plots of variation for other values of j is identical to that shown here, for $j = 9$.

ure 11(B) holds for treated and control groups, i.e., the average mediating effect does not depend on treatment status and hence is the same for all Scratchers. For an alternate interpretation, Figure 11(A) shows the estimated values and confidence intervals for the proportion of *ATE(j)* that is mediated via creation of new projects, given by $\frac{\bar{\delta}^j(1)}{\bar{\delta}^j(1)+\bar{\xi}^j(1)}$. Almost $40-50\%$ of the effect of peers' popularity on Scratchers' future production popularity is explained by the creation of new projects in all future periods subsequent to the treatment at t.

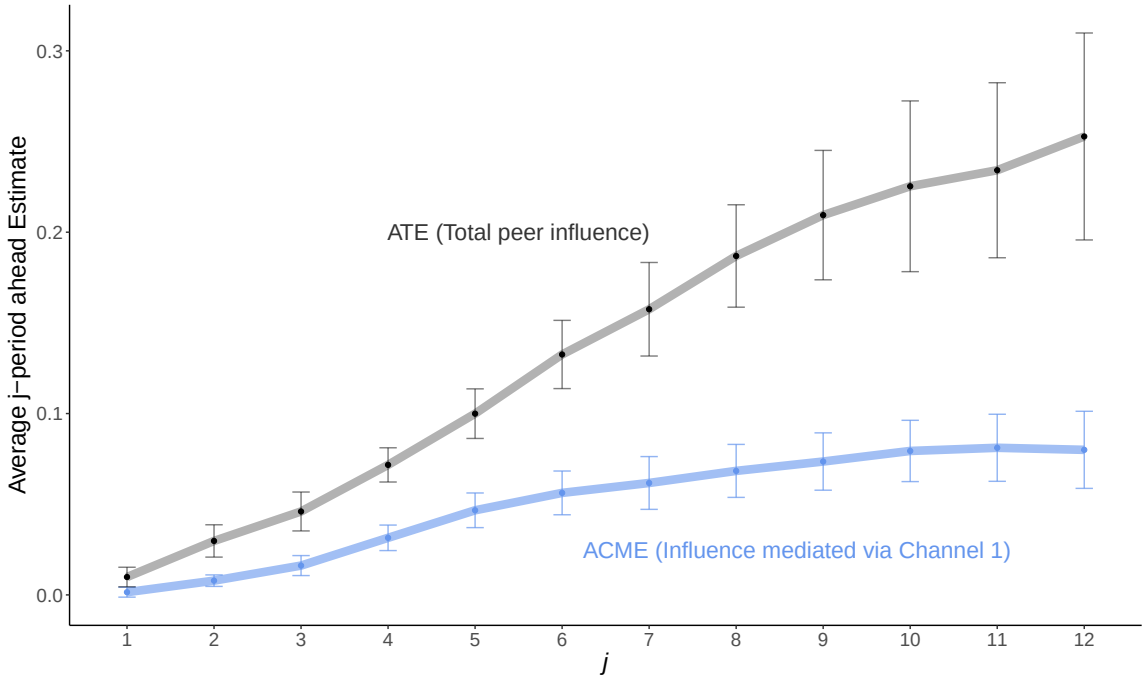
We perform robustness check for the estimates obtained in Figure 11. Identification of estimates $\bar{\delta}^j$ and $\bar{\xi}^j$ assumes sequential ignorability, a set of following assumptions: (i) exogeneity of Tr^t in models (1) and (3) conditional on X^t, N^t, and (ii) exogeneity of $\Delta M^{t\to t+j}$ conditional on Tr^t and X^t, N^t. Assumption (ii) can be violated, even if Tr^t is randomly assigned, by post-treatment variables that affect both mediating and outcome variables in (4). Since this is an untestable assumption, a sensitivity analysis [36] is used to gauge the reliability of the estimates $\bar{\delta}^j$ and $\bar{\xi}^j$ in Figure 11. For this ρ^j in (5) is used as the sensitivity parameter

since variables that violate assumption (ii) will be present in both models (3) and (4). The estimates in Figure 11 are obtained by assuming sequential ignorability, i.e., $\rho = 0$. Sensitivity of *ACME(j)* estimate to other values of ρ is shown in Figure 12 for $j = 9$; for other values of j the shape and values are same as that of the sensitivity curve shown in Figure 12. It would require a high degree of unconfoundedness to violate sequential ignorability in our sample because the estimates of *ACME(j)* will turn 0 only when ρ is 0.6, which is a very high value.

There is an alternative interpretation [38] for the sensitivity test: ρ^2 is the product of unexplained R^2 values of models (3) and (4). So the value of ρ^{j^2} for which *ACME(j)* estimates will turn 0 is 0.36, i.e., *ACME(j)* estimates in Fig. 11 can be refuted if ρ^{j^2} in our sample is 0.36. Let us consider the case of $j = 1$. In our estimated models, the R^2 values for (3) and (4) are 0.028 and 0.43 respectively. $R^2 = 0.051$ in model (1) (Table 5(b), Model 2) increases to $R^2 = 0.43$ in (4) when the mediating variable is introduced for analysis of the causal pathway of treatment effect. To obtain a product value of 0.36, for instance, with a confounder explaining about 60% of the unexplained R^2 of 0.972 in (3), it still needs to explain about 0.62 (108%) in (4) which is well above the maximum R^2 value of 100%. So based on these sensitivity tests, we can not claim that the sequential ignorability assumption is violated in our empirical analysis. The estimates obtained in Figure 11 are therefore valid, and provide a causal interpretation of the mediation of peer influence in production popularity via the creation of new projects by Scratchers.

Finally we present the persistence of the peer influence channel. For this we estimate *ACME(j)* curve for each of the last six months of 2010. The average values of the *ATE(j)* curves and *ACME(j)* curves during this period are shown in Figure 13. This plot confirms that creating new projects in future forms an important channel due to which the production popularity of Scratchers increases in future periods, in response to existing production popularity of peers.

Figure 13: Persistence of Peer Influence Channel



ATE: The average j-period ahead peer influence effect during July-Dec 2010. This curve is the same as in Figure 10 (c). ACME: The average j-period ahead peer influence effect mediated via Channel 1 (creation of new projects) during July-Dec 2010. Error bars are scaled standard deviations of j-period ahead influence estimates.

6 Conclusion & Discussion

We analyzed peer influence in the feature-rich Scratch community under a quasi-experimental setting. Our method accounts for peer influence after controlling for various other mechanisms that can lead to clustering of behaviour in friendship network, i.e., Scratchers and their peers having similar or dissimilar behaviours. We saw how peer influence affects behaviours (creation of new projects, consuming similar projects) and outcomes (production popularity) in the collective learning environment, Scratch. We found a persistent effect of peers' production popularity on Scratchers' future production popularity, and that a large proportion of this effect is mediated by Scratchers' decision to create new projects. We also saw that users who specialize in creating "remixed" projects are susceptible to peers' production popularity more in the short run than in the long run. For consumption behaviour, we found that Scratchers are not influenced by their peers. In particular, we saw that the tendency of users to consume projects from specific sources can not be attributed due to influence from their peers. Although the primary aim of this study has been to understand the role of peer influence in the production and consumption of projects on the Scratch platform, the study has also provided several methodological insights that can inform future work. In retrospect, we believe our results are surprising, especially because production popularity of peers (resulting from the likes of several other users) influences self-decision making to create new projects which would not have been created in absence of such popularity of peers' projects. New projects created due to peer influence attain higher popularity in future compared to projects which are created by users who do not follow "popular friends". (The Scratch platform that provides such a collective sharing and exchange of ideas on a digital platform is therefore valuable because peer influence may not exist if users were to create projects without such a wide exposure to others' projects.) Our study contributes to the literature on production and consumption behaviour [39, 40, 41, 42, 43] (including peer influence [44, 45]) in various knowledge sharing platforms [46, 47, 48, 3]. We believe our results are relevant for a broad audience including network science researchers and practitioners and designers of future educational platforms [49, 50, 51, 52].

6.1 Limitations stemming from data

It would be definitely better if we could have data on several more factors to ensure a higher reduction in bias. (Unfortunately, such fine details are not available in the dataset we have.) So here we discuss some potential limitations of our analysis from a context of the data available to conduct the analysis. *First* is the representativeness of the sample, i.e., we only analyze users who actually joined the platform and interacted in some ways depending on how well they liked the platform under situations existing at the time of their joining. Since the study includes users coming from various countries, we expect that the results hold true in general. *Second*, comparing changes in future behaviour across treated and control groups does not eliminate bias arising out of unobserved confounders that are heterogeneous across the groups – for instance, it may be the case that the general ability of users to operate on the platform might be more in the treated group, and so they are able to find better peers and also can produce more projects. Such confounding effects arising out of ability are however low in our study because we found a significant persistence effect when Scratchers were matched on all their personal characteristics (Table 4(a)) – since this includes various attributes, we expect it also captures the ability of Scratchers which is unobserved. Also, we do not have a priori reasons or information to predict why unobserved variables (like ability) might be distributed differently across experimental and control group, especially in presence of a well balanced matched sample. *Third*, there is an implicit assumption that a Scratcher follows another user in order to be informed about the user's future activities. However, Scratchers happen to follow users for other unobserved reasons as well. Such reasons include help received in a project, social contact in real life, received friendly comment on a project, and joining a particular gallery [53]. Although we account for selection mechanisms in network formation by including peers' observed confounding attributes, our estimates do not control for unobserved selection processes as mentioned above. *Fourth*, we do not have additional data (e.g., survey data [54]) to know exactly the decision making process of Scratchers. Our assumption about a Markov nature of decision making was motivated by an intuition of decision making in large social networks in general (which has also been used in several studies concerning social networks [11, 55, 56]). Although additional data might have been helpful to validate such an assumption, we believe this assumption is not too strong. The assumption used for this study is a weak assumption in the sense that although it

does not capture the trend of past behaviour, it captures a summary of the trend (the aggregate count) which we believe is reasonable. This is because such aggregates (over the entire history) are the only statistics available when a Scratcher browses another Scratcher's profile before following, and when new users who join the platform see about others' projects and activities, and decide their future activities. [5] Given the vastness of the users and projects, and the complexity of the existence of several interactions on Scratch platform, we believe that it is reasonable to assume that the average population (mostly composed of young children) decides its future activities based on the current state of activities and not by the heterogeneous trends in the past leading up to the current state of users' network and projects characteristics. (We have previously mentioned the plausibility of such an assumption in the context of Scratch platform in Section 4.)

6.2 Validity and interpretation of results

We saw that exact matching produces extremely low bias in treatment assignment compared to propensity score matching [20] which, if employed as a tool for analysis, would require a more careful inferential analysis [57, 58, 59][6], especially in presence of many features or variables describing the platform (users, projects, users network, various interactions). We believe that the peer influence estimates have low degree of bias. Although this comes at a cost of more than 50% reduction in sample size (refer Tables 3 and 5), we expect that there should not be an abrupt loss of generality of the results when speaking about the entire population of Scratchers. This is because matching procedure has been performed at different time periods to produce persistence curves as shown in Fig. 10 (c) and Fig. 13. Despite the fact that the users who are dropped out of analysis due to constraints of exact matching are random and not in control of the researchers, these curves are quite smooth and have similar patterns and estimated values as in Fig. 10 (b)

5. A strong Markov assumption, on the other hand, would mean that future decisions are made using information (personal and peers attributes) of aggregate activities from the current month alone. However a weaker assumption allows for future decision to be based not only on the current month's statistics, but also on all previous months' activities summarized as an aggregated counts.

6. Although a balance of propensity scores is necessary for removing selection bias [20], it is not sufficient – the confounders should also be balanced across treated and control groups.

and Fig. 11 (B) respectively. Hence we believe the results are true for the entire population of Scratchers at large.

We believe the description of exact matching in Section 4 is sufficient for the purpose of our analysis since we achieved both a reasonable balance of covariates (and much better than propensity score matching), as shown in Table 4, and a reasonable number of observations for statistical estimation and hypothesis tests of significance of peer influence effects, as shown in Table 5. However, we would like to provide additional details for interested readers to explain better the role of using exact matching in causal estimation of the peer influence parameters [34]. We need to note that matching is not an estimation method. The essence of estimation strategy used in this study is to compare future changes of Scratchers with peers having higher degree of behaviour (experimental group) to those Scratchers whose peers have lower degree of behaviour (control group) in a way which can be argued to be of experimental standards even if we only have observational data. To achieve this goal, exact matching has been used as a data preprocessing step and has helped us in several ways. First, it helped us to create the experimental group when the treatment (peers' variable) was continuous. This was achieved by dichotomizing the treatment by thresholds, a recommended practice [34], which were also shown in Fig. 9 to not influence the nature of our conclusions. Second, due to its intrinsic property, the exact matching helped us to create groups to mimic experimental standards by achieving balance of covariates. Since exact matching on large number of variables generates loss in data, we always ensured to focus on matching the most significant confounders first. Understanding which variables might be important confounders in the data is done by analysing columns (3) and (4) in Table 3 to understand selection into treatment. Unmatched or variables with poor balance in matching were always included as a part of estimation method during regression analysis. Third, it helped us to remove model dependence from our analysis [34], i.e., the peer influence estimates are not extremely dependent on the variables chosen for regression analysis. This is the reason of presenting two different models in Fig. 8 - the estimates are stable. However we use the dataset matched on both personal and network variables for later analysis because this has the best balance of covariates. Lastly, we would like to mention that "exact matching" does not mean having "exactly similar observations" in the control group for each observation in the experimental group. (In fact, such a situation would be impossible, especially in the space of high

dimensional features.) The resulting dataset produced after exact matching has the property that all variables representing personal or peers' characteristics of an ego have the same observed empirical distribution across experimental and control groups. This is in line with the true meaning of exact matching and the requirements for avoiding selection bias on observable variables [34]. We would encourage readers to interpret the estimated values of peer influence as upper bounds, to allow for decrease in estimates due to potential (unknown) unobservables which might be distributed unevenly across experimental and control groups.

Identification of peer influence in our empirical strategy solely relies on the non-existence of confounding latent variables. As mentioned above in discussing our limitations stemming from data availability, we do not claim absence of unobserved variables.[7] If Scratchers and their peers exhibit homophily on a latent covariate and this covariate is correlated both with peers' behaviour (e.g., high or low degree of peers' production popularity) and ego's future change in behaviour, only then such a variable is a confounder, conceptually. This is because only in this case one can claim that the change in future behaviour of ego was driven by common shocks from the latent variable and not due to peers' behaviour. We believe that such confounding variables are least likely in our analysis. *First*, covariates included in analysis, as shown in Table 2 reflect individual preferences for producing projects (e.g., total projects and remixes), individual preferences for consuming projects (e.g., favorites, comments), collective preference of platform users to consume an individual's projects (e.g., love-its, downloads, favorites, comments), attributes reflecting general statistics of platform usage (e.g., age, activity), peers' attributes of all individual properties mentioned above, and characteristics of local network. These variables already reflect a wide range of individual preferences for behaviour on the platform to create projects and build peers network. Activity statistics, which is an important representation of how a user understands the platform details, are well balanced for individuals and their peers. *Second*, as shown in and Fig. 12, the sensitivity test demands very high values of ρ (0.6) for violation of sequential ignorability assumption required for mediation analysis conducted in Section 5. Given that the variables are well

7. Peer influence estimates can become less biased if we could have data on further specific details that reflect individual preferences. However, we believe that the current dataset already has details of a wide range of features that summarize well the activities on the platform.

balanced (Table 4), we believe that the likelihood of existence of confounding variables in our analysis, enough to violate sequential ignorability assumption, is very remote. Since variables that can violate sequential ignorability assumption are also the ones that pose threat to peer influence estimates, we believe that latent variables which are of confounding nature are a least likely case in our study.

6.3 Behaviours analysed in this study

While there are several production and consumption behaviours that may be analysed on Scratch platform, our choices of production popularity and consumption specificity were guided by the following reasons. (We encourage investigation of other behaviours in Scratch, and also in other platforms, in future.) *First*, we wanted to ensure that the behaviours we investigate are plausibly widely known in the Scratch community. For example, a Scratcher knows about his peers' production popularity (e.g., when somebody loves one of his peers' projects) and consumption patterns (e.g., when one of his peers consumes project by favoriting it) through activity feeds. Although a user may not assimilate everything that shows up on activity feeds in real time, we believe that a general knowledge of repetitive behaviour of such peers' activities over a certain duration of time might influence the user to adopt similar behaviour. *Second*, it seems that popularity is a factor that affects social behaviour in general (outside Scratch) [60]. Popularity may be indicative to users who are aware about the activities on the Scratch platform about the popularity-to-quality ratio, i.e., popularity of a project might be indicative of the project content (e.g., codes, creativity, etc.) and hence other users might be interested to learn such things. In this sense, project popularity on the Scratch platform may be seen as a form of collective assessment of the project and thus an increase in a user's production popularity may correspond to an improvement in his/her (unobserved) ability to create better projects. So production popularity may not be as irrational, as a factor to generate influence on self-decisions, as it may otherwise seem to an individual not using the platform. (In retrospect, we indeed find users' behaviour being influenced by peers' production popularity.) *Third*, while peers' projects may influence a user's behaviour, the user might be influenced to a certain extent to consume projects similar to his peers. While tracking each project for each peer is an extremely unlikely situation,

we believe that a consumption influence might exist if a Scratcher observes that most of his users tend to consume a "certain group" of projects. While there may exist several ways to identify such groups of projects, we believe our strategy is feasible on a large scale [61] and also conveys important meaning. We categorized consumption baskets/groups in an intuitive fashion (analogous to various products in a supermarket): projects that are consumed together by most users were placed in one category. (In doing so, all projects have been included in one of the communities and there is no loss of observations.) Later we found from our analysis that users do not tend to be influenced if their peers have high specificity for such a consumption source/category. In retrospect, we investigated and have clearly stated that such communities do not correspond to themes or topics of projects, and neither the choice of network algorithm to detect communities affect our findings. We also believe in retrospection that the inability to differentiate such communities by a particular attribute can be a potential reason why no peer influence exists for consumption patterns. In fact, if the consumptions baskets are largely similar to each other, the reasons for switching to peers' consumption patterns is minimal. (We believe that specificity of consumption of projects is potentially a result of the local network to which a user gets associated to during his/her joining to the platform and formation of initial local friendship network.) In any case, our choice of analysis of consumption behaviour was largely guided by general intuition rather than alignment with forward-looking results. *Fourth*, a user could have hundreds of peers whom he follows but it seems implausible to be influenced by each of them is a heterogeneous and meaningful way. Therefore we used the aggregate measurements of each attribute as a potential source of influence. Later, as shown in Table 1, we found that actually there is no effect of variance of peers' popularity on how users are influenced. In other words, the influence from production popularity is largely an aggregate effect from the popularity of all projects from all peers of a user and not an effect arising from specific peers.

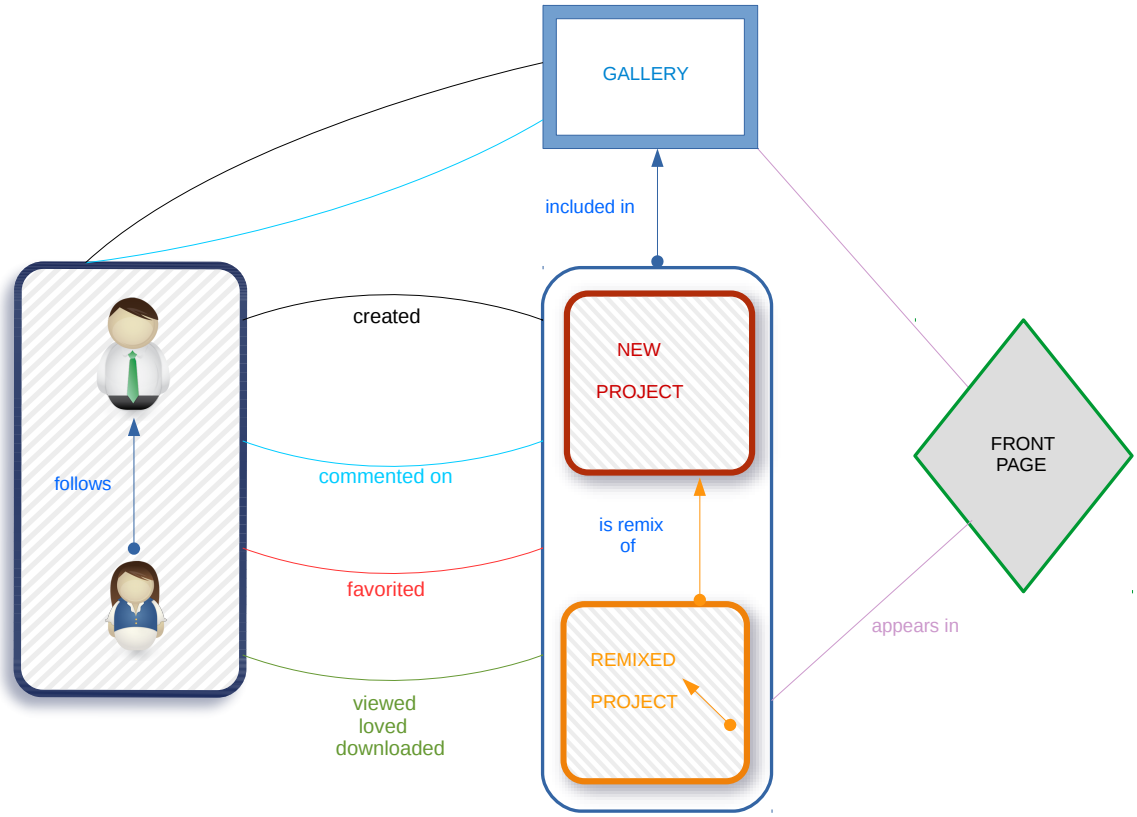
6.4 Some topics for further investigation

We discuss several studies that may be done in the Scratch platform and other digital platforms. *First* is to understand the aspect of "learning" more precisely. In this study, we saw that Scratchers decide to create projects (new, remixed) in

future due to peers' influence; some of the new projects may be totally copied versions without being assigned as remixed projects. So due to the nature of the available data, we can not be precise about how much Scratchers actually 'learn' during the process of creating new projects. *Second*, analysis may be done using other assumptions about users' behaviour which can lead to new insights. Such assumptions can be identified from surveys, or observed behaviours on other platforms. *Third*, for peer influence analysis, especially with many attributes as in this study, ways to reduce dimensionality of the attribute space and their effects on the bias of influence estimates may be conducted. *Fourth*, new studies may be done to better understand differences in peer influence in digital contexts (as in Scratch) and physical contexts (as in classrooms). A key difference between online platforms and classrooms in formal educational systems is that, in most cases, children do not choose to go to schools whereas they usually choose whether to join a platform. Peer influence investigation in physical settings shows mixed evidences [2, 3, 62] of positive and negative influences. New studies can therefore bring clarity into subtle nuances of how children in educational environments are influenced by their peers.

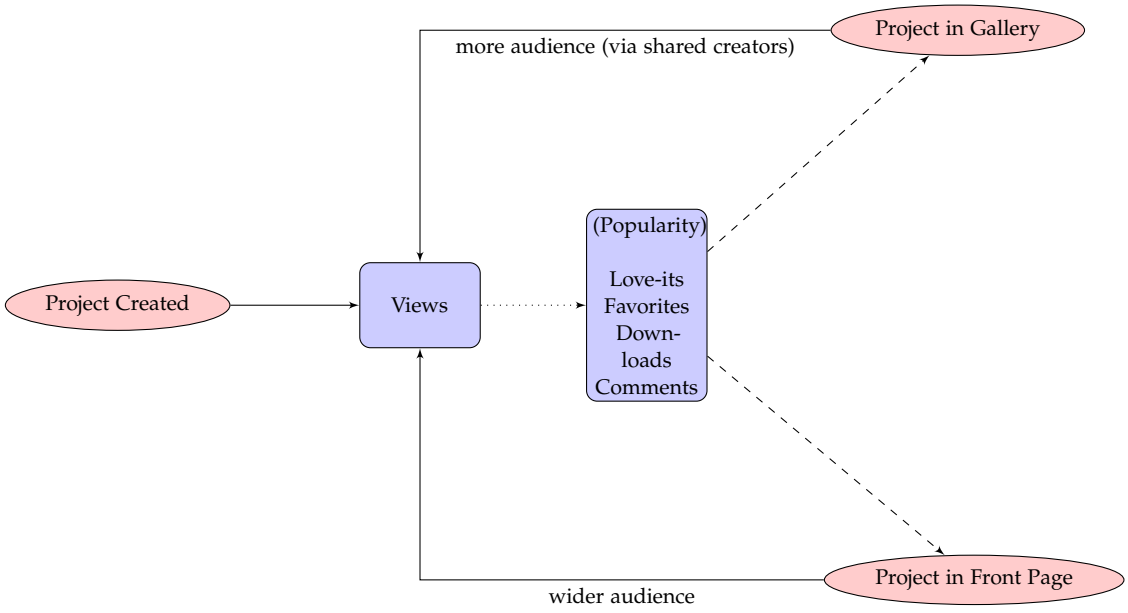
A Appendix

Figure 14: The Scratch Platform



Users produce projects by creating and sharing on the platform. Projects are of two types – new and remixed. Users consume projects by commenting, favoriting, viewing, loving, and downloading them. Views, love-its, and downloads are anonymous. Users can follow each other, and form a friendship network. Projects can be included in galleries, created by shared users. Projects and galleries can be selected to appear on the front page.

Figure 15: Accumulation of Love-Its



The pathways leading to accumulation of love-its. After creation and sharing of a project, love-its on the project can accumulate from views by (1) followers of its creator, (2) users who view it when it appears on the front page (after it becomes popular due to various factors), (3) followers of shared creators of a gallery where the project appears, and (3) random views on the project due to users' browsing.

Table 2: Variables Description

Short Name	Desciption of Variable[†]
Love-its	Total love-its received on all projects created upto t
Views	Total views on all projects created upto t
Projects	Total projects created upto t
Galleries	Total galleries created upto t
Remixes	Total remixes (among all projects) created upto t
Age	Prior to t, number of months in which ego interacted on the platform*
Active	True if ego interacted* more than one month, prior to t
Favorited By	Total projects (of others) favorited by the ego (as a consumer) upto t
Comments (P) By	Total comments made by ego on (own and others') projects upto t
Comments (G) By	Total comments by ego on galleries upto t
Is Remixed	Of all projects created upto t, total projects which have been remixed at least once (at any time)
Projects in Galleries	Total projects (of ego) appearances in various galleries created as of t
Front Page Projects	Total projects that appeared in front page upto t
Featured Galleries	Total galleries created by ego that were featured on the front page upto t
Studio Galleries	Total galleries of ego that appeared in studio design section upto t
Downloads	Total downloads (by others) of projects created by ego upto t
Favorites On	Total favorites (by others) of projects created by ego upto t
Comments On	Total comments received on projects created upto t
Featured Projects	Total projects that were featured on the front page upto t
Following	Total users the ego is following as of t
Followers	Total users who follow the ego as of t
Reciprocation	Total users who follow ego and are also followed by ego as of t
Peers Love–its	Total love–its received by all peers' projects upto t
Peers Views	Total views received by all peers' projects upto t
Peers Projects	Total projects created by all peers upto t
Peers Galleries	Total galleries created by all peers upto t
Peers Remixes	Total remixed projects created by all peers upto t
Peers Active	Total peers who have interacted* more than one month prior to t
Peers Fav By	Total favorites clicked by all peers upto t
Peers Is Remixed	Total projects of all peers upto t which have been remixed at least once (at any time)
Peers Proj in Gall	Total projects (by all peers) appearances in various galleries upto t
Peers Fpage Proj	Total projects created by all peers which appeared in front page as of t
Peers Following	Total users all peers are following as of t
Peers Followers	Total users who follow any peer of the ego as of t

([†]) t refers to a given point in time
(*) Recorded forms of interactions only; does not include views, love-its, downloads because these interactions are anonymous. So if an ego stayed on the platform for only 1 month and downloaded many projects, his age evaluated at any future time is 0.

Table 3: Model Variables, Confounders

Objective	Determine Model Variables		Determine Confounders	
Dependent variable:	Change in Production Popularity†		Treatment $(1/0)^{\dagger\dagger}$	
	[OLS Regression]		*[Logistic Regression]*	
	(1)	(2)	(3)	(4)
Peers Love-its (Tr^t)	0.341^{***}	0.296^{***}		
Love-its	0.077^{***}	0.075^{***}	0.011^{***}	0.018^{***}
Views	-0.003^{***}	-0.003^{***}	0.0001	-0.001^{***}
Projects	0.021^{***}	0.022^{***}	-0.007^{***}	0.007^{***}
Galleries	-0.061^{***}	-0.061^{***}	0.144^{***}	-0.051^{*}
Remixes	-0.022^{***}	-0.029^{***}	0.020^{***}	-0.001
Age	-0.024^{***}	-0.010^{***}	0.021^{***}	-0.005^{*}
Active	0.353^{***}	0.128^{*}	0.549^{***}	0.150^{**}
Favorited By	0.011^{***}	0.007^{***}	0.026^{***}	0.008^{***}
Comments (P) By	0.003^{***}	0.003^{***}	0.013^{***}	0.003^{***}
Comments (G) By	-0.0004^{***}	-0.0002^{***}	0.0001	-0.0002
Is Remixed	0.021^{***}	0.020^{***}	-0.006^{**}	-0.009
Projects in Galleries	-0.0005	0.001		
Front Page Projects	0.286^{***}	0.304^{***}	-0.061^{***}	-0.047
Following	0.005^{***}	-0.041^{***}	0.120^{***}	-0.096^{***}
Followers	-0.026^{***}	-0.020^{***}	-0.009^{***}	-0.008^{*}
Reciprocation	0.190^{***}	0.237^{***}	-0.695^{***}	-0.186^{***}
Peers Views		0.00002^{***}		0.004^{***}
Peers Projects		-0.001^{***}		-0.010^{***}
Peers Galleries		0.001		
Peers Remixes		-0.0002		
Peers Active		0.081^{***}		0.140^{***}
Peers Fav By		0.001^{***}		0.003^{***}
Peers Is Remixed		0.0001^{***}		-0.007^{***}
Peers Proj in Gall		-0.0003^{***}		0.002^{***}
Peers Fpage Proj		0.007^{***}		-0.008
Peers Following		-0.0001^{***}		0.00005
Peers Followers		-0.001^{***}		-0.001^{***}
Constant	-0.115^{*}	-0.106	-1.165^{***}	-4.381^{***}
Observations	73,510	73,510	73,510	73,510
R^2	0.296	0.307		
Adjusted R^2	0.296	0.307		
Log Likelihood			$-39{,}693.690$	$-7{,}285.876$
Akaike Inf. Crit.			79,419.380	14,621.750
Residual Std. Error	7.365	7.304		
F Statistic	$1{,}815.397^{***}$	$1{,}164.492^{***}$		

Notes: *p<0.1; **p<0.05; ***p<0.01
This table shows regression results for $t =$ Dec 2010, $j = 1$.
† Production Popularity: Love-its, †† Treatment: Peers Love-its

Table 4: Balance of Covariates

	Raw Sample	P.Score Matching	(*a*) Exact Matching (X)	(*b*) Exact Matching (X,N)
Love-its	17.65	17.65	0	0
Views	364.69	365.25	0.03	-3.23
Projects	13.42	13.42	0	-0.93
Galleries	0.8	0.8	0	0.01
Remixes	4.39	4.39	0	-0.24
Age	6.4	6.41	0	0.69
Active	0.22	0.22	0	-0.06
Favorited By	11.49	11.49	0	0.29
Comments (P) By	101.68	101.68	0	1.13
Comments (G) By	42.69	42.69	0.12	0.89
Is Remixed	3.69	3.7	0	-0.15
Projects in Galleries	12.11	12.12	-0.02	-0.11
Front Page Projects	0.27	0.27	0	-0.03
Following	17.04	17.09	0	-1.36
Followers	14.43	14.47	0	-1.09
Reciprocation	-0.08	-0.08	0	-0.13
Peers Views	107508.14	107507.79	16051.33	1757.93
Peers Projects	1454.85	1454.91	168.59	-6.92
Peers Galleries	68.34	68.41	7.1	0.12
Peers Remixes	453.48	453.48	52.26	-5.18
Peers Active	17.08	17.09	0.56	-1
Peers Fav By	1508.62	1508.61	160.76	18.2
Peers Is Remixed	1241.27	1241.28	163.41	0
Peers Proj in Gall	2108.5	2108.5	300.93	62.99
Peers Fpage Proj	68.73	68.73	9.59	0.93
Peers Following	2568.72	2569.78	296.73	0
Peers Followers	3326.23	3327.05	406.33	50.86
Featured Galleries	0	0	0	0
Studio Galleries	0	0	0	0
Downloads	51.56	51.66	0	-0.55
Favorites On	12.77	12.78	0	-0.05
Comments On	103.69	103.73	0.05	-0.19
Featured Projects	0.02	0.02	0	0
Treated Group	36697	36697	4502	2380
Control Group	36813	36697	11873	17570
Total Obs.	73510	73394	16375	19950

First panel contains variables used during regressions. Second panel contains variables that were not used for analysis due to one of these reasons: (i) correlated to love–its and carry similar information of a project's popularity, or (ii) multicolinearity detected during regression. Third panel contains count statistics. Columns (in order) show the balance of covariates for $t =$ Dec 2010 before matching, after propensity score matching, after exact matching using X–type variables only, and after exact matching using X– and N– type variables. In (a), all variables are used for matching and in (b) only a subset of all variables is used (Exact matching on N–type variables is expensive.). Balance of an attribute is the difference of (weighted) means of the attribute between treated and control groups; weights are created in (a) and (b) when one user in treated group is matched to many users in control group.

Table 5: Peer/Network Effect

Dependent variable:	Change in Popularity (Love-its) of Projects			
Matched sample used:	(a) Matched on X		(b) Matched on X, N	
	(Model 1)	(Model 2)	(Model 1)	(Model 2)
Peers Love-its (Tr^t)	0.017^{***}	0.017^{***}	0.037^{***}	0.032^{***}
Love-its		0.013		0.040^{***}
Views		0.0003		-0.0003^{*}
Projects		-0.001		-0.001
Galleries		0.001		-0.029^{***}
Remixes		0.031^{***}		0.014^{***}
Age		-0.002^{***}		-0.003^{***}
Active		0.022^{***}		0.035^{***}
Favorited By		-0.001		-0.003^{***}
Comments (P) By		0.003^{***}		0.005^{***}
Comments (G) By		-0.0001		0.0002
Is Remixed		-0.032		-0.002
Projects in Galleries		-0.011		0.004^{***}
Front Page Projects		0.013		-0.012^{*}
Following		0.002		0.002
Followers		-0.004^{**}		-0.001
Reciprocation		0.016^{***}		0.010
Peers Views	0	0	-0.00001^{***}	-0.00000
Peers Projects	-0.00001	-0.00001	-0.0001	-0.0001
Peers Galleries	-0.0001	-0.0001	0.004^{***}	0.004^{***}
Peers Remixes	-0.00000	-0.00000	0.0002	0.0002
Peers Active	-0.0004	-0.001	0.002	0.0003
Peers Fav By	0	0	0.0002^{***}	0.0002^{***}
Peers Is Remixed	-0.00000	-0.00000	0.001	0.001
Peers Proj in Gall	-0.00001	-0.00001	0.0001^{***}	0.0001^{**}
Peers Fpage Proj	-0.00003	-0.00004	-0.003^{**}	-0.002
Peers Following	-0.00000	-0.00000	-0.0002^{***}	-0.0001
Peers Followers	-0.00001	-0.00000	-0.0001^{**}	-0.0001^{***}
Constant	0.004	-0.005	0.018^{***}	-0.021^{***}
Observations	16,375	16,375	19,950	19,950
R^2	0.001	0.005	0.004	0.052
Adjusted R^2	0.0003	0.004	0.004	0.051
Residual Std. Error	0.205	0.204	0.393	0.383
F Statistic	1.363	3.057^{***}	7.255^{***}	38.969^{***}

Notes: *p<0.1; **p<0.05; ***p<0.01
This table shows regression results for t = Dec 2010, $j = 1$.

B References

[1] Winter Mason and Duncan J. Watts. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, 2012. doi: 10.1073/pnas.1110069108.

[2] Jan Feld and Ulf Zölitz. Understanding peer effects - on the nature, estimation and channels of peer effects. Technical Report ROA-RM-2016/1, Maastricht University, 2016.

[3] Bruce Sacerdote. Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics*, 116(2):681–704, 2001.

[4] Rabbany Reihaneh, Samira Elatia, Mansoureh Takaffoli, and Osmar R. Zaïane. *Educational Data Mining: Applications and Trends*, chapter Collaborative Learning of Students in Online Discussion Forums: A Social Network Analysis Perspective, pages 441–466. Springer International Publishing, 2014. doi: 10.1007/978-3-319-02738-8_16.

[5] Benjamin Mako Hill and Andrés Monroy-Hernández. A longitudinal dataset of five years of public activity in the scratch online community. *Scientific Data*, 4(170002), 2017. doi: 10.1038/sdata.2017.2.

[6] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415): 295–298, 2012. doi: 10.1038/nature11421.

[7] Sinan Aral and Dylan Walker. Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science*, 60(6):1352–1370, 2014. doi: 10.1287/mnsc.2014.1936.

[8] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014. doi: 10.1073/pnas.1320040111.

[9] Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010. doi: 10.1126/science.1185231.

[10] Charles F. Manski. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542, 1993.

[11] Tom A.B. Snijders, Gerhard G. van de Bunt, and Christian E.G. Steglich. Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1):44–60, 2010.

[12] Tom A.B. Snijders. Stochastic actor-oriented models for network dynamics. *Annual Review of Statistics and Its Application*, 4(1):343–63, 2017. doi: 10.1146/annurev-statistics-060116-054035.

[13] Christian Steglich, Tom A. B. Snijders, and Michael Pearson. Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*, 40(1):329–393, 2010.

[14] Tom A.B. Snijders. Statistical models for social networks. *Annual Review of Sociology*, 11(37):131–53, 2011. doi: 10.1146/annurev.soc.012809.102709.

[15] Kevin Lewis, Marco Gonzalez, and Jason Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1):68–72, 2012. doi: 10.1073/pnas.1109739109.

[16] Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, 2009.

[17] Bryan S. Graham. Methods of identification in social networks. Technical Report 20414, NBER Working Paper, 2014.

[18] Charles F. Manski. Identification problems in the social sciences. *Sociological Methodology*, 23(1):1–56, 1993.

[19] Camila F. S. Campos, Shaun Hargreaves Heap, and Fernanda Leite Lopez de Leon. The political influence of peer groups: experimental evidence in the classroom. *Oxford Economic Papers*, 69(4):963–985, 2017. doi: 10.1093/oep/gpw065.

[20] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009. doi: https://doi.org/10.1073/pnas.0908800106.

[21] Wikipedia. Scratch. https://en.wikipedia.org/wiki/Scratch_(programming_language), May 2018. [Last accessed 31-05-2018].

[22] Scratch-Wiki. Activity feeds. https://en.scratch-wiki.info/wiki/Activity_Feeds, May 2018. [Last accessed 31-05-2018].

[23] Scratch-Wiki. Project copying. https://wiki.scratch.mit.edu/wiki/Project_Copying, May 2018. [Last accessed 31-05-2018].

[24] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. URL http://stacks.iop.org/1742-5468/2008/i=10/a=P10008.

[25] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, December 2004. URL https://link.aps.org/doi/10.1103/PhysRevE.70.066111.

[26] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2): 026126, 2003. doi: 10.1103/PhysRevE.67.026126.

[27] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE*, 9(6):e98679, 2014. doi: 10.1371/journal.pone.0098679.

[28] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, C. Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *ACM Transaction on the Web*, 6(2):9:1–9:33, 2012. doi: 10.1145/2180861.2180866.

[29] Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011. doi: 10.1177/0049124111404820.

[30] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001. doi: 10.1146/annurev.soc.27.1.415.

[31] Maureen T. Hallinan and Richard A. Williams. Students' characteristics and the peer-influence process. *Sociology of Education*, 63(2):122–132, 1990. doi: 10.2307/2112858.

[32] Robert Huckfeldt and John Sprague. Networks in context: The social flow of political information. *The American Political Science Review*, 81(4):1197–1216, 1987. doi: 10.2307/1962585.

[33] Matthew D. Atkinson and Anthony Fowler. Social capital and voter turnout: Evidence from saint's day fiestas in mexico. *British Journal of Political Science*, 44(1):41–59, 2014. doi: 10.1017/S0007123412000713.

[34] Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, 2007. URL http://gking.harvard.edu/files/abs/matchp-abs.shtml.

[35] Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart. Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8), 2011. URL http://gking.harvard.edu/matchit/.

[36] Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334, 2010.

[37] Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. mediation: R package for causal mediation analysis. *Journal of Statistical Software, Articles*, 59(5):1–38, 2014. doi: 10.18637/jss.v059.i05.

[38] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1): 51–71, 2010.

[39] Jenny Davis. Prosuming identity: The production and consumption of transableism on transabled.org. *American Behavioral Scientist*, 56(4):596–617, 2012. doi: 10.1177/0002764211429361.

[40] Andrea Forte, Judd Antin, Shaowen Bardzell, Leigh Honeywell, John Riedl, and Sarah Stierch. Some of all human knowledge: Gender and participation in peer production. In *Proceedings of the ACM 2012 Conference on Com-*

puter Supported Cooperative Work Companion, CSCW '12, pages 33–36, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1051-2. doi: 10.1145/2141512.2141530. URL http://doi.acm.org/10.1145/2141512.2141530.

[41] Julia Adams and Hannah Brückner. Wikipedia, sociology, and the promise and pitfalls of big data. *Big Data & Society*, 2(2):2053951715614332, 2015. doi: 10.1177/2053951715614332.

[42] Csilla Rudas, Olivér Surányi, Taha Yasseri, and János Török. Understanding and coping with extremism in an online collaborative environment: A data-driven modeling. *PLOS ONE*, 12(3):1–16, 03 2017. doi: 10.1371/journal.pone.0173561.

[43] Gerardo Iñiguez, János Török, Taha Yasseri, Kimmo Kaski, and János Kertész. Modeling social dynamics in a collaborative environment. *EPJ Data Science*, 3(1):7, Sep 2014. doi: 10.1140/epjds/s13688-014-0007-z.

[44] Nabeel Gillani, Taha Yasseri, Rebecca Eynon, and Isis Hjorth. Structural limitations of learning in a crowd: communication vulnerability and information diffusion in moocs. *Scientific Reports*, 4(6447), 2015. doi: 10.1038/srep06447.

[45] Lauren E. Sherman, Patricia M. Greenfield, Leanna M. Hernandez, and Mirella Dapretto. Peer influence via instagram: Effects on brain and behavior in adolescence and young adulthood. *Child Development*, 89(1):37–47, 2018. doi: 10.1111/cdev.12838.

[46] Corey Phelps, Ralph Heidl, and Anu Wadhwa. Knowledge, networks, and knowledge networks: A review and research agenda. *Journal of Management*, 38(4):1115–1166, 2012. doi: 10.1177/0149206311432640.

[47] Todd Rogers and Avi Feller. Discouraged by peer excellence: Exposure to exemplary peer performance causes quitting. *Psychological Science*, 27(3):365–374, 2016. doi: 10.1177/0956797615623770.

[48] Dan Davis, Ioana Jivet, René F. Kizilcec, Guanliang Chen, Claudia Hauff, and Geert-Jan Houben. Follow the successful crowd: Raising mooc completion rates through social comparison at scale. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, LAK '17, pages

454–463, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4870-6. doi: 10.1145/3027385.3027411.

[49] Rhona Sharpe and Greg Benfield. The student experience of e-learning in higher education: A review of the literature. *Brookes eJournal of Learning and Teaching*, 1, 01 2005.

[50] Aftab Dean and Andy Lima. Student experience of e-learning tools in he: An integrated learning framework. *European Journal of Social Science Education and Research*, 4(6):39–51, 2017. ISSN 2312-8429. doi: 10.26417/ejser.v11i2.p39-51.

[51] Chia-Wen Tsai. Do students need teacher's initiation in online collaborative learning? *Computers & Education*, 54(4):1137 – 1144, 2010. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2009.10.021.

[52] Arun Sundararajan, Foster Provost, Gal Oestreicher-Singer, and Sinan Aral. Research commentary—information in digital, economic, and social networks. *Information Systems Research*, 24(4):883–905, 2013. doi: 10.1287/isre.1120.0472.

[53] Scratch-Wiki. Friend. https://en.scratch-wiki.info/wiki/Friend, May 2018. [Last accessed 31-05-2018].

[54] Kevin Lewis, Marco Gonzalez, and Jason Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1):68–72, 2012. doi: 10.1073/pnas.1109739109.

[55] Peter J. Carrington, John Scott, and Stanley Wasserman, editors. *Structural Analysis in the Social Sciences*, page 329–329. Structural Analysis in the Social Sciences. Cambridge University Press, 2005. doi: 10.1017/CBO9780511811395.014.

[56] Vikram Krishnamurthy, Omid Namvar Gharehshiran, and Maziyar Hamdi. Interactive sensing and decision making in social networks. *Foundations and Trends® in Signal Processing*, 7(1-2):1–196, 2014. ISSN 1932-8346. doi: 10.1561/2000000048. URL http://dx.doi.org/10.1561/2000000048.

[57] Dean Eckles. Identifying peer influence effects in observational social network data: An evaluation of propensity score methods. Technical report, Stanford University, 2010.

[58] Dean Eckles and Eytan Bakshy. Bias and high-dimensional adjustment in observational studies of peer effects. Technical report, MIT, 2010.

[59] Gary King and Richard Nielsen. Why propensity scores should not be used for matching. Technical report, Harvard University, 2016.

[60] Bonnie Stewart. Open to influence: what counts as academic influence in scholarly networked twitter participation. *Learning, Media and Technology*, 40 (3):287–309, 2015. doi: 10.1080/17439884.2015.1015547.

[61] Ana Lucía Schmidt, Fabiana Zollo, Michela Del Vicario, Alessandro Bessi, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. Anatomy of news consumption on facebook. *Proceedings of the National Academy of Sciences*, 114(12):3035–3039, 2017. doi: 10.1073/pnas.1617052114.

[62] Scott E. Carrell, Bruce I. Sacerdote, and James E. West. From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica*, 81(3):855–882, 2013. doi: 10.3982/ECTA10168.