Online Urbanism: Interest-based Subcultures as Drivers of Informal Learning in an Online Community

Ben U. Gelman, Chris Beckley, Aditya Johri, Carlotta Domeniconi, and Seungwon Yang*
George Mason University, 4400 University Drive, Fairfax, VA 22030, U.S.A.
*Louisiana State University, 267 Coates Hall, Baton Rouge, LA 70803, U.S.A.
{bgelman, cbeckley, johri, cdomenic}@gmu.edu; *seungwonyang@lsu.edu

ABSTRACT
Online communities continue to be an important resource for informal learning. Although many facets of online learning communities have been studied, we have limited understanding of how such communities grow over time to productively engage a large number of learners. In this paper we present a study of a large online community called Scratch which was created to help users learn software programming. We analyzed 5 years of data consisting of 1 million users and their 1.9 million projects. Examination of interactional patterns among highly active members of the community uncovered a markedly temporal dimension to participation. As membership of the Scratch online community grew over time, interest-based subcultures started to emerge. This pattern was uncovered even when clustering was based solely on social network of members. This process, which closely resembles *urbanism* or the growth of physically populated areas, allowed new members to combine their interests with programming.

Author Keywords
Informal Learning; Online Communities; Interest-based Subcultures; Scratch; Programming.

INTRODUCTION
Recent studies suggest that many youth have adopted information and communication technologies as learning tools outside of school. Levin & Arafeh [35], in a national survey of US teens, reported a 'digital disconnect' between students and their schools. Students reported limited use of the Internet in school when compared to how they used the Internet outside of school – finding virtual textbooks, backpacks, study groups, and guidance counselors. In another study of 4000 middle school students in the USA, students recounted their technology-using practices outside of school and suggested that more 'creative,' 'interactive,' and 'media oriented' uses of technology in school would lead to their increased engagement [34]. The debate on the role of technology in schooling is ongoing but there is enough evidence to suggest that learners are increasingly engaging in opportunities outside of school to advance their learning. In many cases learners' motivation might be more collaborative and enjoyment driven, but the use of online communities is increasing. As Ito et al. [31] found from their extensive study of youth and use of technology, "A smaller number of youth also use the online world to explore interests and find information that goes beyond what they have access to at school or in their local community." They argue that participation in online groups enables youth to connect to peers "who share specialized and niche interests of various kinds, whether that is online gaming, creative writing, video editing, or other artistic endeavors." Youth find these activities fulfilling as they are "fuelled by their interest but also a lack of support for interest-based activities in formal education." Ito et al. argue that in these "interest-driven" networks "youth may find new peers outside the boundaries of their local community. They can also find opportunities to publicize and distribute their work to online audiences and to gain new forms of visibility and reputation (pg. 1)."

In this paper we present a study of an interest-based online community, Scratch ™, designed specifically to provide an informal learning environment for developing programming skills [22]. The research question that motivated our study was: What structural and interactive patterns define a growing online community? As scholars looking at homophily in social networks have suggested [42], there is a need to understand the "ways in which networks evolve over time through cumulative processes of tie creation and dissolution (pg. 438)." We use clustering to examine how the community grew over time to better understand the role of the community aspects – interactions and connections – on informal learning. We found that as the community expanded it supported the development of interest-based subcultures thereby providing users the ability to develop and align with their identities while creating projects that are meaningful to them. We term this phenomenon "online urbanism". Similar to the rise of urbanism in cities, we found that new participants coalesced around new ideas or created artifacts that reflected new ideas [1].

L@S 2016, April 25-26, 2016, Edinburgh, Scotland Uk
© 2016 ACM. ISBN 978-1-4503-3726-7/16/04…$15.00
DOI: http://dx.doi.org/10.1145/2876034.2876052

LITERATURE REVIEW

Informal Learning

The term "informal learning" is often used as a contrast against didactic teaching in formal environments [36] and constitutes a multitude of activities that are "predominately unstructured, experiential, and non-institutional" [39, p.4] and are likely to take place outside the classroom [38]. Other characteristics associated with informal learning include unplanned and incidental cognition which occurs in daily routines [38] or intentional cognition with a particular emphasis on collaborative knowledge building [37]. Bell et al [28] argue that informal environments are shaped by individuals' cultural-historical backgrounds and prior knowledge, identity, and interests are especially important in informal environments, "where opportunities to learn can be fleeting, episodic, and strongly learner-driven" [28, p. 297]. And, "informal environments may particularly foster capacities that are unlikely to register traceable effects on conventional academic measures, notably around interest and motivation and identity" [28, p. 304]. Informal learning requires higher levels of self-regulation and a stronger sense of belonging and identity and often learners have to rely on the community.

Online Learning Communities

Online communities have become a core foundation of how the Internet and the Web are used for a range of issues from health and wellbeing to disaster management. They have also emerged as a popular mechanism for supporting informal learning [22, 40], either through communities designed directly to support learning (e.g. CSILE) [37] or indirectly through efforts such as Citizen Science [28], discussion forums or Q&A sites [44]. Given the overall popularity of online communities, their formation, growth, and decline of online communities have been areas of great interest and there is robust literature around motivation to participate, design of online interactions to increase participation and contributions, the problems faced by communities as they grow, and the need for a critical mass of participants for sustenance [43]. In spite of numerous research studies on this topic, we have limited understanding of structural changes over time, particularly when a community is going through a growth period. This is an interesting issue to examine if we want to better understand what drives the growth of online learning communities – what makes participants both stick around as well as what activities can attract new participants – so that we can design and sustain more productive online learning communities [40].

RESEARCH CONTEXT: SCRATCH ONLINE COMMUNITY

Online communities take many different forms such as discussion forums, question and answer (Q&A) communities, those where knowledge sharing and helping behaviors are critical, or communities such as Wikipedia, where user contributions are essential for the growth of the community. The research context for this study, Scratch, has been characterized by Sylvan as an Online Community of Creators (OCOC) [13]. OCOCs, according to her, are social network sites where the core activity of members is sharing personal and original creations [13]. This characterization of Scratch as an OCOC is of particular relevance within the context of this study. A community of creators offers its participants some unique ways to contribute that differ from other popular online communities such as discussion forums or Q&A sites and sites that only allow sharing of content, for instance photo sharing sites like Flickr. In addition to question and response and commenting on content, an OCOC allows participants to remix and appropriate content based primarily on the interest of the contributors. This process is organic and changes as participants with new interests join

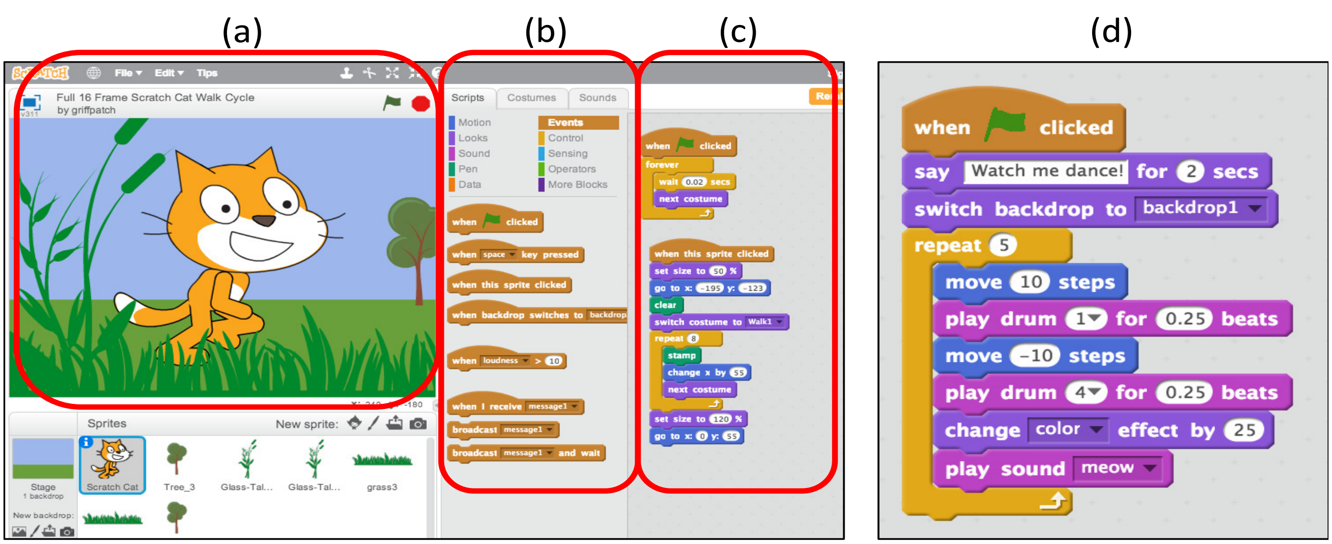

Figure 1. Scratch interface: (a) a project window, (b) blocks, (c) scripts, and (d) a script constructed with blocks.

the community. Also, unlike sites such as Wikipedia where all edits are monitored, Scratch is not a moderated platform.

Scratch is a programming language designed to teach programming basics through a visual, user-friendly interface. The primary target audience is people between the ages of 8 and 16 but the language is open to all [10, 23]. A core goal of Scratch is to make the creation and sharing of interactive content easy for users. As shown in Figure 1, the Scratch programming environment enables users to create programs with blocks, which map to programming constructs and allow users to manipulate data, including media content (e.g., sounds, videos, and images). Users can then share these interactive creations on the Scratch website [24]. The website is an online community where users may run and build on others' ideas and projects. Members can import the source code of any project into their own workspace and remix it. The Scratch website permits a user to view the code of any shared projects. Users may also extend an existing project and make it their own by "remixing". Each user creates a profile on the Scratch website where he shares his creations. Other users can show their interest by tagging projects using tags such as "love" or "favorite" or comment on these projects. Users may also choose to follow others, which allows them to see tags and project uploads of users they follow. Finally, users can chat, discuss project ideas, and work collaboratively in "studios."

In addition to its goal of creating and sharing programs, a fundamental element behind its creation has been to provide an alternate model for how the Web can be used for learning. The Scratch team envisioned learning outcomes related to programming and math skills as well as skills in design, creativity, communication, and so on. Learning is inherently a collaborative process that occurs across communities of practice. The Scratch community was created with the goal to foster learning among users through interaction and sharing of ideas and projects. Resnick et al. wrote about the Scratch in 'Programming for all' and their goal has always been to develop an approach to programming that would appeal to people who had not considered themselves programmers [19]. In May 2007, a website was launched to allow for more sharing of projects on the Web. Maloney et al. studied the use of Scratch at a Computer Clubhouse where urban youth between the age of 8-18 used Scratch [14]. The findings from this research indicated that users learned key programming concepts even in the absence of instructional interventions or experienced mentors. The ease of use for the Scratch platform and its interactivity made it attractive for users.

Data Description and Selection

The Lifelong Kindergarten Group at the MIT Media Lab created the Scratch dataset from the Scratch Online Community [24, 25]. The data set contains 1,056,950 registered users and 1,928,699 projects created between March 5, 2007 and April 1, 2012. Demographics of users are not published due to privacy concerns considering that the majority of Scratch users are kids and teens. This dataset was made publicly available by the Scratch team [26]. The number of users in the community shows a steep linear increase starting from around 100,000 users in March 2008 to slightly over one million users in March 2012 (Figure 2). The attributes associated with users include follower/followed connections, join dates, and number of projects. Project attributes include creation dates, types of code blocks used to create the project, and whether the project was remixed from others. The number of followers and the number of projects are highly skewed, with the majority of the population making 0 or 1 follower connection and/or project (see Figures 3 and 4 respectively). Of the total number of users, 143,234 users (13.55% of total users) have at least one follower, and 304,357 users (28.80% of total users) have created and shared at least one project. Overall, 66.68% of all users (704,820) has never shared a project or followed other users. This long-tail phenomenon is similar to user activity in other online communities [4].

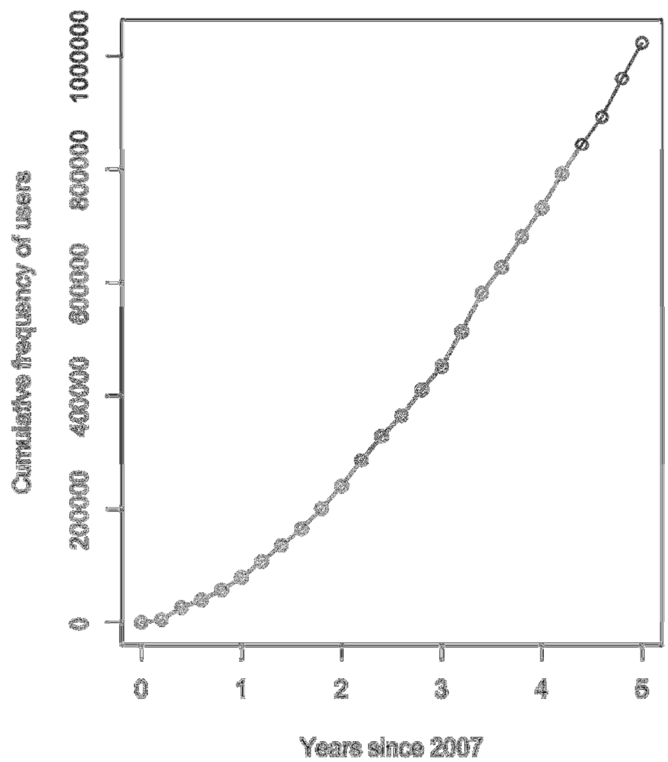

Figure 2: The growth of Scratch users ('0' in the x-axis represents March 5, 2007, and '5' represents March 5, 2012).

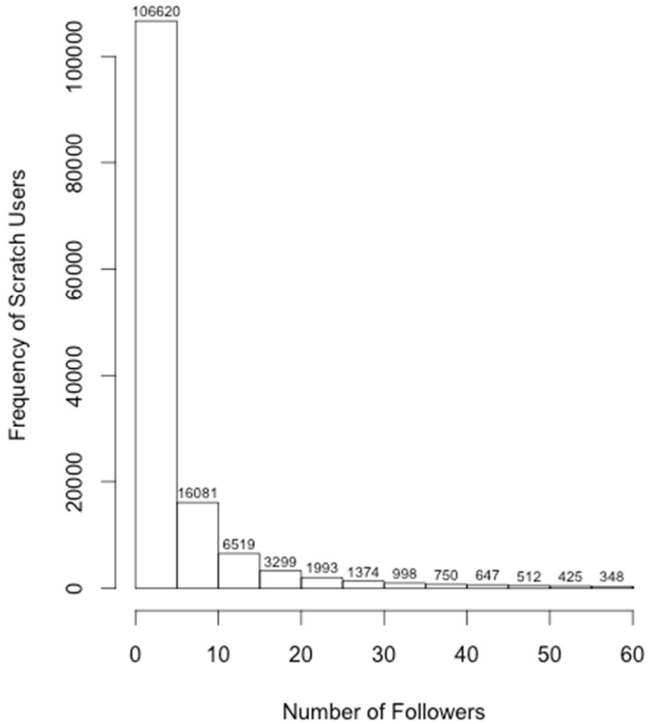

Figure 3: Frequency of Scratch users by the number of followers.

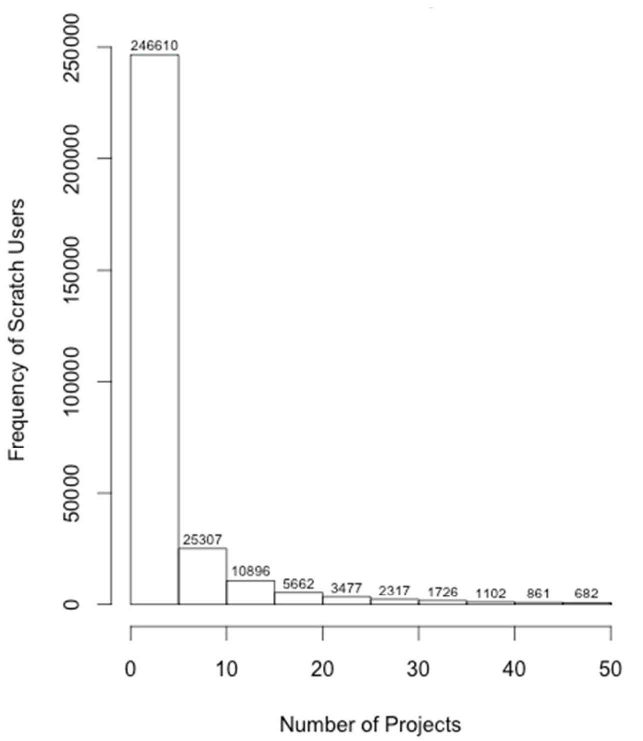

Figure 4: Frequency of Scratch users by the number of projects.

Follower-Followed Relationship as a Measure of Social Ties in Scratch

As recommended by [21], the investigation of social ties, beyond kinship, is critical for understanding subculture formation. Therefore, to select core active members in the community we leveraged the "follower/followed" relationships between pairs of users, rather than the number of projects a user has produced. In fact, the number of followers was a better indication of a user's activity: a user may attract followers if s/he not only is highly active, but also creates projects of interest to others. To generate a list of core active contributors, we adapted the concept of k-core of a graph introduced in [6]. Given the directed graph representing the follower relationship between users, we selected the users with at least k followers (i.e. the nodes with in-degree larger than k). The selection was performed once and not recursively as in [6]. As a result, our notion of k-core is less restrictive: for example, if a node n has in-degree k, and some of the followers have in-degree less than k, n is still considered as a k-core user.

Number of followers	Number of users
1+	95,780
10+	18,835
25+	8,154
50+	4,228
75+	2,738
100+	1,971

Table 1: Cumulative user counts with varying followers.

As shown in Table 1, we investigated different cutoff values for k and eventually settled on k = 25. We observed that at higher thresholds there was a slight bias towards long-term users as they had more time to accumulate followers. For instance, a cutoff value of k = 100 resulted in interactions and project themes of 1,971 users. In contrast, a cutoff value of 25 resulted in 8,154 active users including those who joined the community recently. Considering that the users with 25+ followers gave us a more representative sample, we performed in-depth analysis for these 8,154 users and their 468,057 follower/followed relationships.

Analysis: Identification of Subgroups

In recent years, several studies have examined different models of community growth over time to understand how members' behavior and dynamics impact community participation [4]. In this study, we used clustering techniques to examine growth in community participation. We used highly active core members as proxies for community activity as their participation significantly shapes communities' persistence and growth particularly during periods of high growth [2]. To achieve a better understanding of the collaborative nature of online communities scholars have focused primarily on structural features such as the underlying social networks within the communities. Within this tradition, multiple quantitative methods have been proposed for identifying the structure of a community. For instance, [15] has used hierarchical clustering whereas [16] used a behavioral method to classify members. Researchers have also used a 'sense of community' instrument [17] and [18] used a k-plex analysis to identify sub-communities as cohesive subgroups within a community.

OpenOrd Clustering

The OpenOrd layout in Gephi [5] was used to map the structure. OpenOrd is a multi-level, force-directed layout and uses average-link clustering based on both edge weights (if present) and distance, where distance is determined using a force-directed algorithm [8]. Clusters of nodes are replaced by single nodes, and the clustering is repeated until a certain distance threshold between the nodes is reached. After the clustering is complete, the graph is expanded by replacing the individual nodes with the original graphs in each cluster [8].

Clustering of Users with 25+ Followers

We clustered 8,154 users who have at least 25 followers, and identified five large clusters as shown in Figure 5. Cluster A was the oldest of the three – in terms of the average join dates of its members – and heavily featured early Scratch developers from MIT. Other users in the cluster developed complex projects that stretched the limits of Scratch as a programming language. Cluster B was younger on average than cluster A, and users began to focus more heavily on making games. Cluster C was the youngest of these three clusters, and contained users who had a large variety of projects. Cluster D was the largest and the main project theme of this cluster was games. Various types of game projects were identified – from simple ones involving small number of characters and limited use of media, to complex ones having multiple levels of difficulty and advanced multimedia support. Cluster E was the second smallest one and its users were mainly focused on art projects. Especially, the canine and feline images and characters were often appearing in the projects. We present the details of the temporal analysis for A-E in the next section.

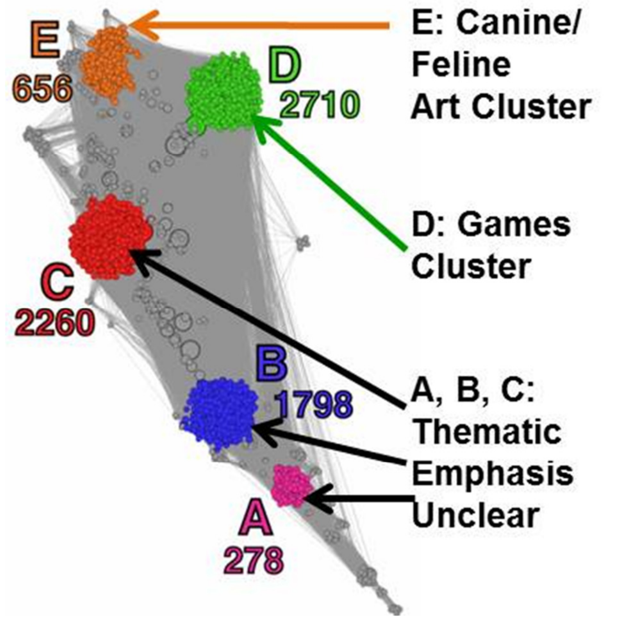

Figure 5: Users with 25+ followers. The number under each cluster label represents the size of the cluster.

To further investigate clusters A-E and uncover similarities within the cluster, we used a text mining approach (Table 2). By concatenating project titles, descriptions, and tags for all projects within each cluster, we represented each cluster as a document with a bag-of-words approach. For each term we calculated the term frequency * inverse document frequency (TFIDF) value [11]. The TFIDF allows uncovering discriminant terms by favoring frequent terms within a cluster, at the same time, penalizing common terms across clusters [11]. Prior to the analysis, all terms were stemmed using the Porter stemmer. The top frequent terms in Cluster A did not indicate any trend. For clusters B and C, certain commonality emerged among the top words. Terms such as "mario," "game," "pokemon," and "sonic" (as in Sonic the Hedgehog™) were present suggesting a slight emphasis on gaming. Cluster D showed a significant increase in the uniformity of the top terms. "Super Mario," "sonic," and "press" (as in "press these keys to play the game") constituted a larger percentage of the total terms as

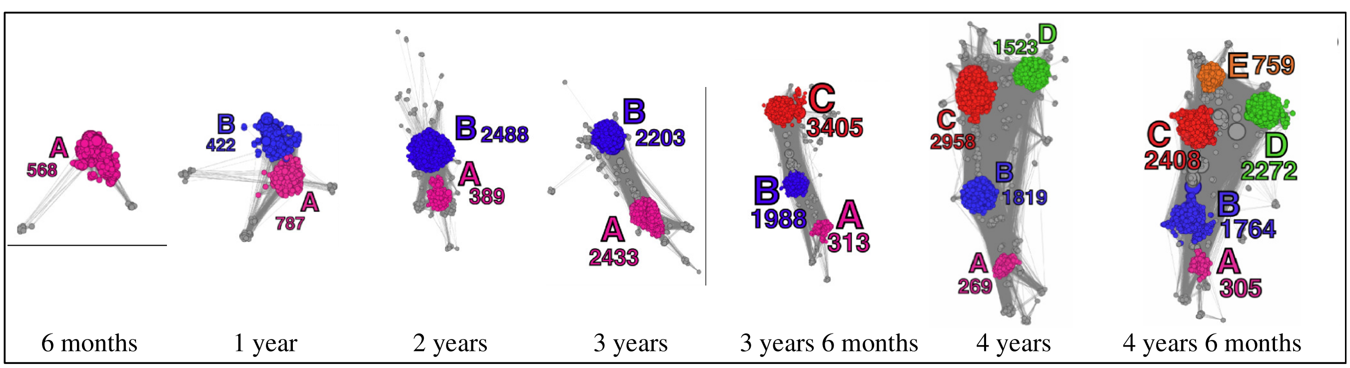

Figure 6: Cluster evolution: Cumulative time of the slicing is marked for each network graph.

compared to clusters B and C, suggesting specialization on gaming within cluster D. Cluster E exhibited top terms that were considerably different from other clusters but the terms had a high degree of uniformity.

In order to examine this cluster further, we selected a sample of the projects from the cluster and inspected them manually. We found that the projects heavily featured canine and feline art. The terms "art" and "contest" (as in a drawing contest) support these findings. Thus, the top terms for cluster E demonstrate a thematic uniformity of feline/canine art. Another reason D and E are interest-driven is the fact that the interest-specific terms also make up a higher percentage of the total terms in those clusters. Although the thematic emphasis of clusters was evident, we informally observed that there was a temporal characteristic to these clusters as well. To further investigate this aspect empirically, we developed the concepts of "Scratchage" and "Scratchend" (see [14] who use a similar methodology).

Temporal Emergence of Subgroups

Scratchage

Scratchage refers to the age of a user's account, measured in years. A user's Scratchage is relative to the end of our dataset. Thus, users who created their account on the last day of our dataset (April 1, 2012) have the youngest Scratchage of zero, and the users who created their account on the first day of our dataset (March 5, 2007) have the oldest Scratchage of 5.08 years. We found a trend among the clusters where as we progressed from cluster A to cluster E the Scratchage of the users decreased even though we did not use any temporal data for clustering (first column in Table 3). We believe that this occurred because of users tend to follow their contemporaries, rather than inactive users.

Scratchend

To measure and explore inactivity in the Scratch community, we created an attribute called "Scratchend," which is the amount of time in years between the final activity of a user and the end of the dataset (April 1, 2012). Our definition of "final activity" is either the final time a user follows someone else, or the final time he/she shares a project (whichever occurs last). For example, cluster D has a Scratchend of 0.37, meaning that, on average, it has been 0.37 years from the last user activity to the end of the dataset. Users who have reached their Scratchend are not likely to obtain any new followers. If a user is not developing friend connections with newer users, who have been joining at an increasing rate, then OpenOrd layout's repulsion forces will push these older, less connected users away. We see this trend in Figure 6, where the oldest clusters A and B are well separated from the younger clusters C, D, and E. We hypothesize that the users in the older clusters A and B joined Scratch, befriended their contemporaries, and then became inactive over time. Along with the increasing rate of new users, this explains why we see a temporal separation of clusters as well as newer clusters C and D with significantly more users.

Cluster	SA: Avg.	SA: SD	SE: Avg.	SE: SD
A	4.85	0.25	3.80	1.28
B	4.11	0.46	2.43	1.25
C	2.79	0.68	1.19	0.86
D	1.69	0.89	0.37	0.46
E	1.25	0.72	0.23	0.30

Table 3: Summary of Scratchage and Scratchend

Community Evolution

In order to dynamically visualize the evolution of the Scratch community, we used a modified version of Falkowski's overlapping sliding window approach [3]. The method involves slicing our dataset at various points in time, clustering each data slice via OpenOrd, and then comparing clusters across time slices. We began by slicing the dataset at one-year intervals. After clustering and sampling each frame, we added certain half-year intervals to observe incremental changes between large transformations in the graph. The initial text analysis of the clusters that we presented in Table 2 is paramount in tandem with the evolution of the Scratch community. The top terms displayed virtually no thematic uniformity in Cluster A, minor uniformity in clusters B and C, and significant uniformity in clusters D and E. This increase in thematic uniformity of the clusters directly corresponds to the age and growth of the clusters. The clusters in Figure 7 are labeled and colored to show the evolution and migration

Cluster	Top Words (TFIDF Values as percentage of Total Terms)
A	comput (5.3); time (4.7); make (4.4); space (4.4); random (4.0); new (3.8); stupid (3.6); first (3.2)
B	waffle (19.3); mario (17.6); game (12.3); super (9.5); tag (9.3); fun (8.5); pokemon (8.0)
C	funni (16.2); waz (15.0); wuz (15.0); tag (11.8); pokemon (10.4); good (8.6); add (8.6); sonic (8.5)
D	mario (22); epic (20.7); tag (16.3); press (12.2); sonic (9.8); super (8.4); add (6.2); scratch (4.2)
E	warrior (11.5); wolf (9.6); yay (8.8); cat (7.9); contest (7.7); pleas (7.0); art (5.7); pie (5.7)

Table 2: Summary of stemmed top frequent words in each cluster.

of users in the clusters over the five years of the dataset. Initially, all users were clustered into one large community (surrounded by small niche communities). This primordial community contained the majority of users at the time. By the first year, some users began to go inactive, and so the initial cluster, A, started to split in two, as new users arrived and friended the active existing members. This active/inactive separation continued until, directly after the beginning of the third year, cluster B burst and formed a third major cluster, C. Between the years 4 and 4.5, clusters D and E emerged from the enormous C cluster.

DISCUSSION

We draw on the "Subcultural Theory of Urbanism" proposed by Fischer [19]-[20] as framework for interpreting our findings. Within the realm of urbanism, subcultural theory proposes that city life is relatively unconventional (defined broadly) and cities stimulate unconventionality because population concentration generates a variety of subcultures. These subcultures can be seen as a "set of interconnected social networks" with common norms and habits. Unlike other theories of urbanism that argue that city life or urbanism results in breakdowns and thus leads to subgroups, the subculture theory argues that population growth provides the opportunity to find others with whom one can take part in a common way of life through mutual sharing of traits, interests and/or values. *Therefore, in the case of Scratch what we identify through our analysis as subgroups can be better understood as subcultures.* Furthermore, according to Fischer [20], the larger the place the more specialized are the subcultures since the population is higher and therefore there is more cultural heterogeneity. Although heterogeneity might exist in small places as well, the critical mass needed for a subculture is reached only in larger places.

Another characteristic of this theory identified by Fischer is that "subcultural theory is, at core, an ecological theory, not a theory about persons." Therefore, although aggregated individual level data is often used to identify and study urban subcultures, a more ecological approach using techniques such as clustering are appropriate. Subcultures are emergent and rather than *a priori* determination of the unit of analysis, it is more appropriate to let the data determine the unit and "its ultimate tests require place-level analyses and, ideally, emergent, place-level measures [20, pg. 548]." We argue that this theory is applicable in the context of online learning communities since, as Fischer himself contends [20, pg. 549], "subcultural theory is largely about the ability of subculture members to communicate", a prerequisite also for learning. He further argues that "specialized and unconventional aspatial subcultures should emerge for those populations with access to modern means of communication and transportation [20, pg. 550]", which is reflected in the online nature of learning and participation by users from across the world. Overall, according to Fischer, the diversity of urban life actually enhances the number of social ties, in particular those that are voluntaristic in nature. In urban conditions individuals have greater freedom in choosing whom they associate with, other than their kin, ties which are often forced. More than physical proximity, self-selection is the important criteria for subcultures to emerge. From an empirical perspective, to better understand the emergence of subcultures it is important to be able to look at social ties [21] and the concentration of these ties based on cultural affiliation (and not just physical proximity).

Online communities provide an interesting context to test the theory since physical proximity is not usually the cohesive factor in online communities (there are some exceptions), a large community in terms of member size provides an ideal context to test this theory. This theoretical perspective provides a unique lens to examine online communities as participation is largely voluntary and based on some form of affiliation beyond kinship. Our interest in using this framework emerged concurrently with our analysis of the community as we saw similar features – such as interest-based subcultures – emerge in our analysis.

One explanation for the emergence of interest-based

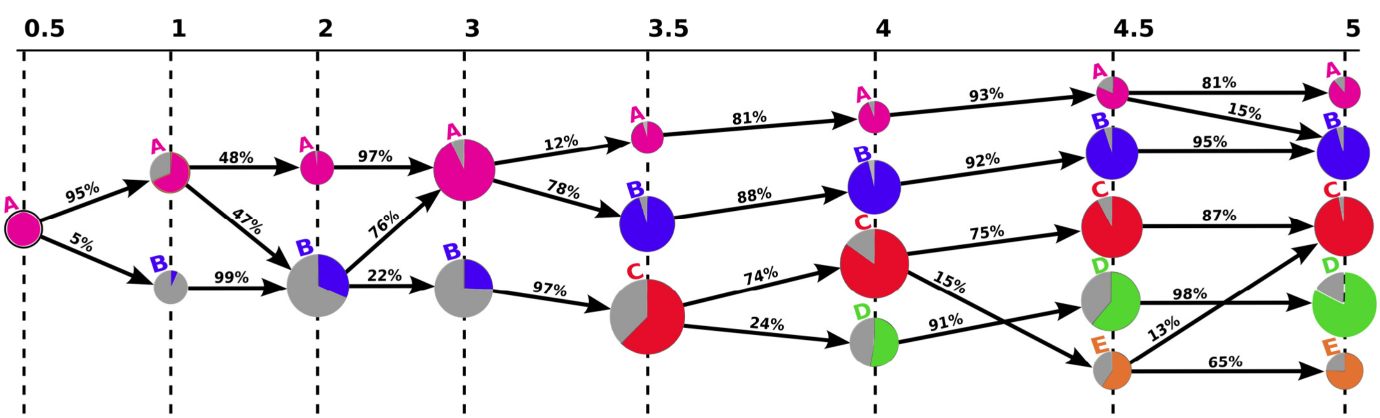

Figure 7: Scratch Community Evolution: Pie-charts' diameters are scaled to represent size of cluster; Light gray indicates new users in each cluster; Colors indicate migrated users in a cluster; Arrows indicate migration (percentage over time).

subculture might be Azevedo's [29] concept of "lines of practice" which suggests that "interests are realized in the specific fabric of activities that a person fashions himself or herself and that by definition extends into many realms of the person's life [29, pg. 505]." So, if learners are engaged with games in their lives, working on designing or programming of games is a natural extension of the "lines of practice." Azevedo [29] further argues that "an interest does not have a core whose essence is topic/domain centered—as when one says, for instance, that a student has an interest in biology, or science, or algebra, or technology, or programming, or physics, and so on, or when one speaks of students developing specific mathematical or scientific identities. Rather, being interested in a practice requires weaving it with many other concerns, domains, values, goals, and practices in one's life space, which makes the practice of interest meaningful in the short and long hauls [29, p. 505]." From this perspective as well we can see how learners create subcultures that allow them to develop an identity not only as a programmer, or someone learning how to program, but also align it with games or animations or with cats. Newcomers to these subcultures are able to see and remix projects from other participants, that is, they can participate peripherally, before contributing complex projects of their own – a core process in learning [45].

In this paper, we show how the Scratch community grew into interest-based subcultures, which act as important building blocks of the community structure that provide benefits to the community members with varying interests. As Azevedo, [29, pg. 464] defines, "interest-based engagement in a practice refers to self-motivated, often self-guided, short- and long-term participation in the fabric of activities that make up the practice." Interest-driven learning is usually differentiated into situational and personal interest [30, 32]. Situational interest refers to a temporary or fleeting engagement often triggered by the task environment. In contrast, an individual or personal interest refers to a person's long-term disposition to engage in a topic or domain. Personal interest and is usually associated with high knowledge and high value of such a topic/domain. The role of online communities in fostering interest-driven learning is consistent with the literature. For instance, Azevedo [41] demonstrated that the level to which students engage with an activity is often highly dependent on infrastructure, physical and digital, that allows students the opportunity to tailor the activities to their interests. This, when combined with the ability to pursue both self-initiated detours and multiple activities, enhances participation and learning [29, pg. 466]. Hidi [33], in her review paper, further emphasizes the importance of interest in learning and development. She argues that interest facilitates learning, improves understanding and stimulates effort as well as personal involvement. This focus on interest was diminished during behaviorism and the subsequent cognitive revolution but has garnered attention as neuroscientists have rediscovered the importance of emotions and feelings and psychologists have started to look at affective variables. Recent findings show positive influences of interest on attention, quantity and quality of learning, and improvement in learners' organization of tasks and their goals and choices. For example, within Scratch, a user can access a large number of projects created by others in the same subculture. Accessing such projects may promote learning of new Scratch blocks through remixing, and eventually enhance sustainability of the community by motivating the user to create similar or more sophisticated projects. The ability to join or create new subcultures provided users with an opportunity to express themselves, to find others with whom they shared interests and ideas, and have their own 'space'. In real world communities whereas subcultures evolve around shared societal goals, similar educational and socio-economic background, shared lifestyle choices, and so on, in Scratch the subcultures evolved around shared thematic interest, in addition to the shared interest in programming.

By examining the subcultures in Scratch, as well as the evolution of the entire community over time, we found a pronounced evolutionary trend. In the beginning, the community existed as a single entity where most active users followed each other and created projects with diverse themes. As the community grew massively and suburbanized over time, many of the initial users became inactive, while others migrated to new subcommunities that attracted new users. Eventually, the newer communities began to diverge from each other and formed distinct entities based upon certain, targeted interests such as games, animations, and arts. Since studies of online communities first appeared, scholars have argued that in many respects the social interactions on them mirror what we find in offline communities [27].

Our study supports as well as extends this view by demonstrating that not only in terms of social interactions at the dyadic or small group level but even looking at larger assemblages and longer term interactions we find that online communities can mirror offline communities. In keeping with the view of the subcultural theory of urbanism [9], we provide an ecological understanding of an online community and show how certain interest areas first thrive and then lose participants as other areas emerge and attract users. This process continues and in time the overall fertility increases providing newcomers with more and more options of subcultures. This work contributes to research on newcomer participation in online communities [43] by demonstrating the strong influence of affinity based participation. With the opportunity to attach themselves to interest-based subgroups, newcomers are able to self-select into subculture that appeal to them [43]. Although we are unable to comment directly on demographic homophily due to the lack of relevant data [42], our findings demonstrate the emergence of interest-based homogeneity at the level of social networks within a larger community.

Implications

There are several implications of this study. When designing online communities, users can be better supported by providing them the opportunity to create interest-based subgroups. Clustering can be used to choose projects that are highlighted in the community. Projects can be introduced strategically to increase visibility of active users from different subcultures. Our analysis highlights both the promises and pitfalls of community dynamics. It suggests that providing users a place to find people like them and to be able to explore their interests is critical in creative communities. On the other hand, this can also lead to divisions but we found a few users acted as brokers and were connected with multiple subcultures.

Limitations

Our study has several limitations:

- Lack of demographic data: This data was not released considering that the majority of Scratch users are minors. Richer analysis of subcultures in Scratch might have been available with this data.
- Selection of $k = 25$: In our analysis, we included users only with 25 or more followers. With $k < 25$, a more detailed analysis for the evolution of the community as well as diverse subcultures may have been possible. With $k > 25$ coupled with user account time stamps, may have allowed us to better understand the subcultures of highly active/long-term participants.
- Cumulative data of follower/followed relationships: Our analysis for the community evolution considers user relationships in a cumulative way. Thus, relationships that are inactive for a while may also have been included in the analysis. However, this cumulative relationship data enabled us to trace the evolution of the community throughout its lifecycle.

Future Work

The primary suggestion for future work is to examine in more detail individual learner's interest development, not just on the Scratch community platform but across their learning ecology. A comprehensive study of this sort will be able to shed light on the overall four-phase model of interest development and provide further guidance on how to better design online communities for learning. It will also allow us to evaluate how to integrate such communities in formal learning environments and understand what Azevedo [29] articulates as "lines of practice".

CONCLUSION

This study sheds new light on the participation of core active members over the lifecycle of an online community. As the population of Scratch grew at a higher rate after it became popular, users split into interest-based groups. Newer users conformed to these groups, thus perpetuating the interest-based communities. Although Scratch has a very specific goal – teaching computer programming – it resembles other OCOCs and online communities in general and these findings are broadly relevant.

ACKNOWLEDGEMENTS

We appreciate the Lifelong Kindergarten group at MIT for publicly sharing the Scratch datasets. This work is partly based upon research supported by U.S. National Science Foundation (NSF) Award#DUE-1444277 & EEC-1408674.

REFERENCES

1. Alba, R., Logan, J., Stults, B., Marzan, G., & Zhang, W. (1999). Immigrant groups in the suburbs: A reexamination of suburbanization and spatial assimilation. *American Sociological Review,* 446-460.
2. Butler, B., et al. (2002). Community effort in online groups: Who does the work and why? *Leadership at a Distance* (2002): 171-194.
3. Falkowski, T. et al. (2006). Mining and Visualizing the Evolution of Subgroups in Social Networks. *IEEE/VIC/ACM Int. Conf. Web Intelligence*.
4. Grabowski, A., Kruszewska, N., & Kosiński, R. A. (2008). Properties of on-line social systems. *European Physical Journal B*, 66(1), 107-113.
5. Jacomy, M., Venturini, T., Heymann, S., Bastian, M. (2014) ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* 9(6): e98679.
6. Kumar, R., Novak, J. & Tomkins, A. (2010). Structure and evolution of online social networks. *Link mining: models, algorithms, and applications*. Springer New York, 337-357.
7. Manning, C., Raghavan, P. & Schütze. H. (2008). *Introduction to information retrieval.* Cambridge University Press.
8. Martin, S. et al. (2011). OpenOrd: An open-source toolbox for large graph layout. *IS&T/SPIE Electronic Imaging*.
9. Massey, D. & Denton, N. (1988). Suburbanization and segregation in US metropolitan areas. *American Journal of Sociology*, 592-626.
10. Resnick, M., et al. (2009). Scratch: programming for all. *Communications of the ACM*, 52(11), 60-67.
11. Robertson, S. (2004). Understanding inverse document frequency: On theoretical argument for IDF. *Journal of Documentation*, 60(5), 503-520.
12. Scratch. http://scratch.mit.edu/about/
13. Sylvan, E. (2010). Predicting Influence in an Online Community of Creators. *CHI 2010*, pp. 1913-1916.
14. Raban, D. R., Moldovan, M., & Jones, Q. (2010). An empirical study of critical mass and online community survival. *Proceedings of CSCW* (pp. 71-80).
15. Hopcroft, J., Khan, O., Kulis, B., & Selman, B. (2004). Tracking evolving communities in large linked networks. *PNAS*, 101, 5249-5253.

16. Chin, A., & Chignell, M. (2006). A social hypertext model for finding community in blogs. *Proceedings of Hypertext and Hypermedia* (pp. 11-22).
17. McMillan, D. W., & Chavis, D. M. (1986). Sense of community: A definition and theory. *Journal of community psychology*, 14(1), 6-23.
18. Chin, A., & Chignell, M. (2007). Identifying active subgroups in online communities. *Proceedings of Center for advanced studies on Collaborative research* (pp. 280-283).
19. Fischer, C. S. (1975). Toward a subcultural theory of urbanism. *American journal of Sociology*, 1319-1341.
20. Fischer, C. S. (1995). The subcultural theory of urbanism: A twentieth-year assessment. *American Journal of Sociology*, 543-577.
21. White, K. J. C., & Guest, A. M. (2003). Community lost or transformed? Urbanization and social ties. *City & Community*, 2(3), 239-259.
22. Yang, S., Domeniconi, C., Revelle, M., Sweeney, M., Gelman, B. U., Beckley, C., & Johri, A. (2015). Uncovering Trajectories of Informal Learning in Large Online Communities of Creators. *Proceedings of Learning @ Scale* (pp. 131-140). ACM.
23. Maloney, J., Peppler, K., Kafai, Y.B., Resnick, M., Rusk, N. (2008). Programming by Choice: Urban Youth Learning Programming with Scratch. *Proc. Of SIGCSE*, 367-371. New York, NY: ACM Press.
24. Scratch™ Online. Retrieved: http://scratch.mit.edu.
25. Scratch™ Online Statistics. Retrieved from http://scratch.mit.edu/statistics/.
26. Scratch Research Data Sharing Agreement. Retrieved from http://llk.media.mit.edu/scratch-data/.
27. Wellman, B., & Gulia, M. (1999). Net surfers don't ride alone: Virtual communities as communities. *Networks in the global village*, 331-366.
28. Bell, P., Lewenstein, B., Shouse, A. W., & Feder, M. A. (2009). *Learning science in informal environments: People, places, and pursuits*. National Academies Press.
29. Azevedo, F. S. (2013). The tailored practice of hobbies and its implication for the design of interest-driven learning environments. *Journal of the Learning Sciences*, 22(3), 462-510.
30. Renninger, A., Hidi, S., & Krapp, A. (2014). *The role of interest in learning and development.* Psychology Press.
31. Ito, M., Horst, H., Bittanti, M., Boyd, D., Herr-Stephenson, B., Lange, P. G., ... & Robinson, L. (2008). *Living and Learning with New Media: Summary of Findings from the Digital Youth Project*. John D. and Catherine T. MacArthur Foundation.
32. Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational psychologist*, 41(2), 111-127.
33. Hidi, S. (2006). Interest: A unique motivational variable. *Educational Research Review*, 1(2), 69-82.
34. Spires, H. A., Lee, J. K., Turner, K. A., & Johnson, J. (2008). Having our say: Middle grade student perspectives on school, technologies, and academic engagement. *Journal of Research on Technology in Education*, 40(4), 497-515.
35. Levin, D., & Arafeh, S. (2002). The digital disconnect: The widening gap between Internet-savvy students and their schools. *Pew Internet & American Life Project*. http://www.pewinternet.org/reports/toc.asp?Report=67/
36. Eraut, M. (2000). Non-formal learning and tacit knowledge in professional work. *British Journal of Educational Psychology*, 70(1), 113-136.
37. Bereiter, C., and Scardamalia, M. (1989). Intentional learning as a goal of instruction (Bereiter & Scardamalia, 1989). In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 361-392). Hillsdale, NM: Lawrence Erlbaum.
38. Hodkinson, P., Colley, H., & Malcolm, J. (2003). The interrelationships between informal and formal learning. *Journal of Workplace Learning*, 15, 113-318.
39. Marsick, V. J., & Volpe, M. (1999). The nature of and need for informal learning. In V. J. Marsick & M. Volpe (Eds.), *Advances in developing human resources: Informal learning on the job* (pp. 1–9). Baton Rouge, LA: AHRD.
40. Barab, S., Kling, R., & Gray, J. H. (2004). *Designing for virtual communities in the service of learning*. Cambridge University Press.
41. Azevedo, R. (2007). Understanding the complex nature of self-regulatory processes in learning with computer-based learning environments: An introduction. *Metacognition and Learning*, 2(2-3), 57-65.
42. McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 415-444.
43. Kraut, R. E. & Resnick, P. (2012). *Building successful online communities: Evidence-based social design*. Cambridge, MA: MIT Press.
44. Teo, H. J. & Johri, A. (2014). Fast, functional, and fitting: expert response dynamics and response quality in an online newcomer help forum. *Proceedings of CSCW 2014*, pp.332-341.
45. Lave, J. & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.