An Empirical Study of Critical Mass and Online Community Survival

Daphne Raban
University of Haifa, Israel
draban@gsb.haifa.ac.il

Mihai Moldovan
New Jersey Institute of Technology, USA
mimold@gmail.com

Quentin Jones
New Jersey Institute of Technology, USA
qgjones@acm.org

ABSTRACT
There is general consensus that critical mass at inception ensures the sustained success of online communities. However, no clear understanding of what constitutes such a 'critical mass' exists and too few quantitative studies have been conducted into the relationship between initial online community interaction and its longer term success to draw any conclusions. In this paper we start to address this gap through a large-scale study of the relationship between IRC chat channel survival and initial chat channel community interactions. A sample 282 chat channel births was used for survival analysis which explored the relationship between the overall user activity in each channel at its inception and the channel's life expectancy. Significant relationships were observed between online community lifespan and critical mass measures: 1) message volume, 2) user population heterogeneity and 3) production functions. The results lend support to the *Critical Mass Theory* of collective action.

Author Keywords
Critical Mass, Online Community System Design, Chat, IRC, Synchronous Communication, Computer-Mediated Communication.

ACM Classification Keywords
H.5.3. Group and Organization Interfaces: Collaborative Computing.

General Term
Measurement.

INTRODUCTION
Communication utilizing collaborative computing systems may seem ephemeral, even chaotic, to an external observer. Some interaction environments support groups of users for long periods of time, while others are short-lived and disappear before long. The term *online community* is often used to describe the groups that coalesce over time utilizing a particular collaborative computing system environment (see [8] for an early history of the term). Several scholars argue that online communities evolve in stages (e.g. [2, 10, 11, 16, 22]), and that each stage has distinct characteristics, that must be taken into consideration for community building efforts. In line with this notion is the idea that to successfully advance from the creation / inception stage to maturity requires the gaining of a critical mass of users (e.g. [6, 8, 9, 14, 18, 19]). The concept of critical mass is imaginatively described by Schelling [21]:

> *An atomic pile "goes critical" when a chain reaction of nuclear fission becomes self-sustaining; for an atomic pile, or an atomic bomb, there is some minimum amount of fissionable material that has to be compacted together to keep the reaction from petering out*.

In the case of online communities, what constitutes the critical mass of users is unclear. It cannot simply be a particular number of users, as not all users participate equally, and if it is more than the users themselves, what other factors need to be taken into account? In reaction to this, some researchers have argued that critical mass is so context dependent as to make the issue moot [1, 3]. However, very little empirical work has been done to explore this issue and its possible implications for long-term group survival and system design.

Here we examine the relationship between critical mass, and more generally initial group interactions in online community spaces, and the longer-term survival of the group interactions. The empirical focus is on Internet Relay Chat (IRC) channels. To date, little is known about the initial conditions that lead to the formation of groups in synchronous spaces such as IRC channels, as well as about the subsequent conditions necessary for those groups to evolve and be sustained over longer periods of time. The paper begins with a discussion of *Critical Mass Theory*, which is followed by the presentation of a large-scale study of the relationship between IRC chat channel survival and initial chat channel community interactions. We conclude with a discussion of the implications for theorists and designers of online community systems.

CSCW 2010, February 6–10, 2010, Savannah, Georgia, USA.

CONCEPTIONS OF CRITICAL MASS

Perhaps the earliest mention of critical mass in relation to collaborative computing was made by Licklider in 1968 [11]. He suggested that computer mediated communication was constrained by the necessity for a "*critical mass*" or minimum number of people to be available online for the solving of various problems. This was followed by the observations made by Hiltz and Turoff in the early 1970s regarding computerized conferencing systems that some discourse groups were simply 'too-small' to sustain interactions [5]. Users of these small groups either stopped using the system or engaged in what they referred to as "electronic migration" to larger, more active groups and conferences. As a result, they hypothesized from their small sample that computerized conferences with less than 8 to 12 active users would not have "critical mass" and after a short while, would fail to produce enough new material to justify users' continued use of the system.

Palme [17] expanded Hiltz and Turoff's early work and proposed a different set of numbers for critical mass for different activities. He also proposed a linear 'communication response function' to explain the group size threshold for sustainable computer-mediated communication (CMC). However the relationship between user-population and user-contributions is far more complex than a simple linear function. We know from various sources that as online groups grow there is a decrease in the likelihood of individual user's participation in public discourse [7, 13].

Further complicating matters is the fact that the term critical mass in regards to CMC sometimes refers to the chances of gaining a response to a message [18, 20], and sometimes the chances of new material being generated [5, 23]. Perhaps most importantly various researchers (e.g. [1, 3]) argue that the minimum number of users required to sustain ongoing discourse is primarily determined by the social context of discourse, which suggests that it may not be possible to develop a general critical mass model.

From the above, we can see that some authors consider critical mass a species of threshold model, in which a minimum number of contributors is necessary for a certain tipping point to be passed, leading to sustainable cooperation [5]. In the sociology [15] and economics literature [21], critical mass is seen from the more general lens of modeling collective action, particularly public goods production functions.

CRITICAL MASS THEORY

In 1985 Oliver, Marwell and Teixeira put forward the *Critical Mass Theory* [15] which provides a complex theoretical model of the production of the collective actions required for longer term group survival. Extending the work described by Olson [16] and Hardin [4], one of the theory's most important contributions is the argument that a group's level of heterogeneity, together with the shapes of various production functions, defined as relationships between resources contributed by the group and the collective output of that group, can be used to predict the likelihood of longer-term success of the group [15]. The production functions are used to distinguish likelihood of longer-term group success. Based on the shape of the graphs obtained by plotting the number of resources by the amount of group success, the *Critical Mass Theory* described two types of production functions: accelerating and decelerating. In the Method section we provide further detail as to the specific functions that emerged in the current study.

Critical Mass Theory was extended to interactive media by Lynn Markus [12] who explained that in the special case of online interpersonal communication reciprocal interdependence emerges and requires interactivity for system's success. In other words, not only posting is important, but also responding to earlier posts is important. More posts require even more responses. In this case an accelerating production function is required which makes the achievement of a critical mass all the more challenging. Another element pointed out by Markus [12] was the importance of time in assessing the diffusion process through the community [12]. It is important to note that the *Critical Mass Theory* as applied to online communication holds that factors used to quantitatively characterize initial interactions will predict longer-term group success. This contrasts with the position of some CSCW researchers who hold that what constitutes critical mass is dependent on the social context.

Here the driving research question is the ability of *Critical Mass Theory*'s to predict the longer-term sustainability of groups in online synchronous interaction spaces. Specifically, it asks whether it is possible to predict IRC channels' chances of survival by looking at some of the initial starting conditions that numerically characterize the overall activity of the channels, at the trajectories of the channel activity occurring inside over various time intervals in the initial stages of the channels' lives, at the population's level of heterogeneity during various time intervals and at the channels' production functions computed for the same time intervals.

Because the question deals with the timing of events, specifically with the lifespan of chat channels, survival analysis methods are in order. This paper examines whether it is possible to distinguish between the likelihood of IRC channels' survival over time based on variables extracted from the analysis of IRC channel interaction dynamics, on their heterogeneity of population, on their trajectories of activity, and on their production functions, all computed for four different time intervals.

Considering the above, it is hypothesized that the long term survivability of any newly born publicly active channel can be predicted using four categories of factors: (1) the level of channel activity during various initial time intervals; (2) the trajectories of channel activity during various initial time intervals; (3) the heterogeneity of the channel's population during various initial time intervals; and (4) the type of production functions for various initial time intervals.

METHOD

Data was collected for one year, from February 1 2005 until January 31 2006. During this time we collected 176,966,975 messages, posted to 7,365 active chat channels, by 489,562 unique nicknames. Two methods were used to capture data: 1) IRC bots (software agents) were used that continuously monitored all the channel spaces and collected data at specific time intervals; and 2) through open source TCP traffic monitoring software and associated custom written data parsing system. Both these methods were enabled by our hosting and administering a server on the AustNet IRC network. Austnet was chosen for the following reasons: (1) it had a medium size with an average of 4,000 users and 2,500 channels at any time; (2) it was a distributed network, consisting of servers located in the US, Europe, Asia, and Australia; (3) English was the predominant language; and (4) its management committee agreed to link a new server to the network.

Users of the network were informed about the research when logging onto the network but the data collection bots and traffic monitoring tools were not visible to network users and did not result in any detectable change in user behavior. Using our administration privileges and customized software we are confident that we were able to reliably distinguish IRC bots and human users. The data reported in this paper pertains only to human users of the AustNet IRC network. The entire data was anonymized.

This data is ideal for exploring the role of *Critical Mass Theory* because it allows for the collection of data about the birth of a large number of interaction spaces and their possible disappearance during the study period. This is explained in the following.

Data Considerations

To conduct this analysis we need to identify new channels, their birth and possible death during the study period. From the larger data we extracted a subset of IRC channels for analysis that came into existence and were "born" during the month of July 2005 and followed them until January 31, 2006.

To carry out this work, it was first necessary to operationalize the notions of online group birth and death. Because of the current lack of research in this area, no well-known definitions pertaining to these terms currently exist. Consequently, there was a need to clearly define them, from the perspective of this work. We considered a group to be three people, and as we were looking for publicly active groups, we considered a channel to be "born" the first day when it hosted at least three posters who exchanged at least four public messages during the same 20-minute interval; and a channel was considered "dead" if four weeks of non activity passed since the last day that channel hosted at least three posters who exchanged at least four public messages during the same 20-minute interval. A channel was considered to be non-active during a particular day if less than three posters were publicly active in that channel during all 20-minute intervals of that day. The main reason for defining the birth and death of chat-channels in terms of the supported level of activity was because of the interest presented by this level of activity. Our research revealed that channels can easily be created and can exist for long periods of time after their creation without being visited at all before they would eventually disappear. Therefore, a channel's creation day and disappearance day may not be relevant indicators for the actual life of the channel. The definitions provided above are better suited for this research because they examine the life and death of a channel based on the presence or absence of public activity inside that channel, and not based simply on the presence or the absence of the channel itself on the IRC network.

Analysis

In order to explore the long term survivability of publicly interactive IRC channels, all the channels that were born during July 2005 were identified. Then, the lifetime of each channel was computed as the number of days between the birth and the death of that channel. A total of 282 channels were born during that month. Out of those channels, only 8 were still alive, according to the above definition, at the end of the data-collection period (January 31, 2006); the other 274 died at some point during the second half of the year for which data was collected. Our aim being to understand how to distinguish the channels that survived from the channels that did not survive, and what factors predicted overall survival, we used the Cox regression analysis.

Cox regression (sometimes called proportional hazards regression) is a method for investigating the effect of several variables upon the time a specified event takes to happen. In the context of an outcome such as death this is known as Cox regression for survival analysis. The method does not assume any particular "survival model" but it is not truly non-parametric because it does assume that the effects of the predictor variables upon survival are constant over time and are additive in one scale.

Cox regression is used for modeling time-to-event data in the presence of censored cases (censored cases are cases for which the event of interest has not been recorded). In this research, the event of interest was the death of the channels, which was observed for 274 cases. Eight cases were censored – the ones corresponding to the channels that continued to be active after the end of the data-collection period. (Note: all of the channels are censored cases until they die - it is not just the 8 at the end).

However, as opposed to other time-to-event modeling methods such as the Kaplan-Meier survival analysis, the Cox regression allows the inclusion of predictor variables (covariates) in the models. Cox regression will handle the censored cases correctly, and it will provide estimated coefficients for each of the covariates, allowing the assessment of the impact of multiple covariates in the same model.

Four Cox regression models were created, corresponding to four different initial starting conditions time intervals for which the predictors of survivability were computed. These time intervals were: (1) the first two hours of life; (2) the first day of life; (3) the first week of life; and (4) the first two weeks of life.

The objective was to determine whether the survival of channels can be predicted by looking at the initial starting conditions that characterized the overall activity of the channels; at the trajectories of the channel activity occurring inside them; at the level of heterogeneity of the channels' populations; and at the channels' production functions, computed for each of the four time intervals mentioned above.

Table 1 describes the variables entered into each Cox regression model. The number of users, posters, lurkers and messages measured the overall channel activity; the posters trajectory (PT) and messages trajectory (MT) variables measured the trajectories of channel activity; and the poster homogeneity (PosterHG) variable measured the homo/heterogeneity of the channel poster population. The dependent variable was the lifespan of the channels, computed as the number of days between the birth and the death. Each variable in Table 1 was measure in four time periods: the first two hours of activity, the first day, the first week, and the first two weeks.

Variables	Description
Users	Total number of users
Posters	Total number of posters
Lurkers	Total number of lurkers (non-posters)
Messages	Total number of messages
PosterHG	Poster homogeneity
PT	Posters trajectory
MT	Messages trajectory
PF	Type of production function

Table 1. Variables used as predictors for life span of chat channel.

The possible values of the PT and message trajectory MT variables ranged from -1 to 1 and they indicated how the number of posters and the number of messages varied over time, during each of the 4 initial starting conditions time intervals. The value represented the slope of the line. For example, a value of 1 in a channel's poster trajectory measure for its first two hours of life would indicate that the number of posters for that channel continuously increased with every 20-minute interval since the channel's birth. As another example, a value of -1 in a channel's MT variable for its first day of life would indicate that the number of messages for that channel continuously decreased with every hour that passed since the channel's birth.

The PT and MT variables were computed for each channel as the Spearman correlation coefficients between time and the number of posters or messages observed in that channel for each of the four time intervals. Each interval had a different number of data points that were used in computing the correlation coefficients. The first two hours of a channel's life had six data points for which the number of posters and messages were computed, each corresponding to a 20-minute interval. The first day of a channel's life had 24 data points, each corresponding to an hour; the first week of a channel's life had 7 data points; and the first two weeks of a channel's life had 14 data points, each corresponding to a day. The time was expressed as the number of seconds that have elapsed since midnight Coordinated Universal Time of January 1, 1970 until the starting time of the data point interval.

The possible values for the PosterHG variables ranged from 1 to 100 and they indicated how heterogeneous or homogeneous the poster population of a channel was during a particular time interval, with respect to a larger time interval. A channel was considered more homogeneous if its poster population stayed relatively constant as time passed (the same people continued to post), and more heterogeneous if its poster population changed significantly over time (a variety of people posted). The maximum value of 100 indicates a fully homogeneous population, while the minimum value of 1 indicates a population with the highest level of heterogeneity.

The poster homogeneity for the first two hours of life was computed as the percentage value represented by the number of posters present in the channel during this interval reported to the total number of posters that visited the channel during its first day of life. The poster homogeneity for the first day of life was computed as the percentage value represented by the number of posters present in the channel during this interval, reported to the total number of posters that visited the channel during its first week of life. The poster homogeneity for the first week and the first two weeks of life was computed as the percentage value represented by the number of posters present in the channel during those intervals, reported to the total number of posters that visited the channel during its first month of life.

For example, consider a channel that had 3 posters during its first two hours, 10 posters during its first day, 20 posters in its first week, 25 posters during its first two weeks and 30 posters during its first month. In this case, PosterHG2Hrs = 3/10 = 30%, PosterHGFirstDay = 10/20 = 50%, PosterHGFirstWeek = 20/30 = 66% and PosterHGFirstTwoWeeks = 25/30 = 83%. Here, the values of the computed diversity variables show that initially the population was more heterogeneous, but with the passage of time it became more homogeneous.

Production Functions

The production functions were computed based on the definition provided by the *Critical Mass Theory*. The theory defined production functions as the relationships between resources contributed by a group and the collective output of that group. Plotting the number of resources by the amount of group success resulted in a variety of production functions which belonged to two types according to the *Critical Mass Theory*: accelerating and decelerating.

In the case of IRC chat-channels, the number of users present in the channel was considered a surrogate measure for the group resources, while the number of messages was considered a surrogate measure for the amount of group success achieved. Twelve categories of production functions were identified after plotting the number of users by the number of messages, for each of the 282 channels. These twelve categories are presented in Table 2.

Category	Shape of the production function
0	Constant
1	Linear ascending
2	Linear descending
3	Accelerating ascending
4	Decelerating ascending
5	S-shaped ascending
6	Accelerating descending
7	Decelerating descending
8	S-shaped descending
9	Parabola
10	Inverse parabola
11	Variable/Unidentified

Table 2. Categories of production functions found in IRC channels.

For each channel, the shape of the production functions for all the four intervals was determined by plotting the number of users by the number of messages for that channel, using the same data points that were used to compute the trajectory measures described above.

It was observed that some of the production function categories identified in the manner described above were not very common. In order to make the analysis easier and more relevant, the twelve categories of production functions were grouped into four broader types: constant, ascending, descending, and variable. Table 3 describes these four types in terms of the categories they included while Table 4 reports the number of channels characterized by each type of production function during each of the analyzed time intervals.

Type	Included categories
Constant	Constant
Ascending	Linear ascending, accelerating ascending, decelerating ascending, S-shaped ascending
Descending	Linear descending, accelerating descending, decelerating descending, S-shaped descending
Variable	Parabola, inverse parabola, variable/unidentified

Table 3. Broad types of production functions found in IRC channels.

Interval	Const.	Ascen.	Descen.	Variable
1st 2 hours	39	136	58	49
1st day	26	164	14	78
1st week	48	171	17	46
1st 2 weeks	38	166	13	65

Table 4. Number of IRC channels in each broad type of production function per time interval.

The Cox Regression Model

The model-building process took place in two blocks. In the first block, a forward stepwise algorithm was employed and the following variables were entered: the number of users, the number of posters, the number of lurkers, the number of messages, the poster diversity, the posters trajectory, and the messages trajectory. In the second block, the categorical variable used to represent the type of production function was added to the model (see Table 1 for the exact names of the variables used in the four regression models corresponding to each time interval). A separate regression model was run for each time period sampled.

RESULTS

The basic model offered by the Cox regression procedure is the proportional hazards model, which assumes that the time to event and the covariates are related through a particular equation. The hazard function is a measure of the potential for the event to occur at a particular time t, given that the event did not yet occur. Larger values of the hazard function indicate greater potential for the event to occur. The baseline hazard function measures this potential independently of the covariates. The shape of the hazard function over time is defined by the baseline hazard, for all cases. The covariates simply help to determine the overall magnitude of the function.

Eight cases of the total of 282 were censored. These cases represented the channels that did not die. They were not used in the computation of the regression coefficients, but were used in the computation of the baseline hazard.

It can easily be observed that half of the new channels that appeared in July 2005 did not last more than a day and had very few users, posters, lurkers, and messages. It is likely that such channels were created by very small groups of users who decided to get together for short periods of time to discuss something in a more private environment, rather than in the open spaces of other already existing channels. The channels disappeared after those discussions were resolved and the users left. It may also be noted that a vast number of channels had more homogeneous populations, rather than heterogeneous, for all four intervals.

Table 5 reports descriptive statistics for the variables entered into the four Cox regression models.

Variable	Mean	Median	Mode	Range
Lifespan	17	1	1	1-203
UsersFirst2Hrs	9	5	3	3-113
PostersFirst2Hrs	6	4	3	3-36
MessagesFirst2Hrs	123	66	4	4-1001
LurkersFirst2Hrs	4	1	0	0-81
PosterHGFirst2Hrs	86	100	100	8-100
UsersFirstDay	15	6	3	3-293
PostersFirstDay	8	4	3	3-87
MessagesFirstDay	217	96	4	4-2202
LurkersFirstday	8	2	0	0-206
PosterHGFirstDay	82	100	100	8-100
UsersFirstWk	27	10	4	3-683
PostersFirstWk	13	6	4	3-184
MessagesFirstWk	410	132	4	4-6249
LurkersFirstWk	15	3	0	0-499
PosterHGFirstWk	93	100	100	18-100
UsersFirst2Wks	33	11	3	3-799
PostersFirst2Wks	15	6	3	3-217
MessagesFirst2Wks	539	135	4	4-8145
LurkersFirst2Wks	19	5	0	0-582
PosterHGFirst2Wks	92	100	100	25-100

Table 5. Descriptive Statistics for the Variables in the Cox Regression Models

The Cox regression models were statistically significant at $p<.000$ for all four time periods sampled, however, each period was characterized by different predictor variables as reported in Table 6.

Period	Cox χ^2	Predictors	Exp(B)
First2Hrs	21.9**	PosterHG	1.013**
FirstDay	46.9**	Messages	.999**
		PosterHG	1.015**
FirstWeek	63.9**	Messages	.999**
		PosterHG	1.026**
First2Weeks	82.2**	Messages	.999**
		PosterHG	1.034**
		Ascend.PF	.692*
		VariablePF	.623*

**p<.001; *p<.05

Table 6. Results of the Cox Regression Models

Table 6 shows that from all the variables entered into the first block of the regression model, only the poster homogeneity contributed significantly to it in the first two hours of the channels' life. The addition of the type of production function to the model as a categorical variable did not contribute to the model for the first 2 hours. For the first day of activity the poster homogeneity and the number of messages contributed significantly to the Cox regression model, again without a contribution by the production functions. The same results were obtained for the first week of the channels' life. The observation from two weeks of the channels' life shows a contribution by the ascending and variable types of production functions to the regression model.

Exp(B) in Table 6 represents the predicted change in the hazard (lifespan, or, in other words, the death of a channel) for a unit increase in the predictor. In this case, the value of Exp(B) for PosterHGFirst2Hrs means that the channels' death hazard increases by (100% * 1.013) – 100% = 1.3% for each increase of 1% in the homogeneity of the channel poster population. The death hazard for a channel whose diversity measure witnesses a raise of 10% is increased by (100% * (1.013^10)) – 100% = 13.78%. Put simply, the larger the value of the poster homogeneity measure, the higher the risk of death (remember that a homogeneity value of 100 signifies a fully homogenous channel while lower values imply greater heterogeneity).

The value of Exp(B) for PosterHGFirstDay means that the channels' death hazard increases by (100% * 1.015) – 100% = 1.5% for each increase of 1% in the homogeneity of the channel poster population. The value of Exp(B) for MessagesFirstDay means that the channels' death hazard decreases by 100% - (100% * .999) = 0.1% for every new message. The death hazard for a channel which would have 100 more messages in its first day of life would decrease by 100% - (100% * (0.999^100)) = 9.5%. In other words, the larger the value of the poster diversity measure, the higher the risk of death, and the more messages, the lower the risk of death.

The value of Exp(B) for PosterHGFirstWeek means that the channels' death hazard increases by (100% * 1.026) – 100% = 2.6% for each increase of 1% in the homogeneity of the channel poster population. The value of Exp(B) for MessagesFirstWeek means that the channels' death hazard decreases by 100% - (100% * .999) = 0.1% for every new message. Same as before, the larger the value of the poster homogeneity measure, the higher the risk of death, and the more messages, the lower the risk of death.

The value of Exp(B) for PosterHGFirst2Weeks means that the channels' death hazard increases by (100% * 1.034) – 100% = 3.4% for each increase of 1% in the homogeneity of the channel poster population. The value of Exp(B) for MessagesFirst2Weeks means that the channels' death hazard decreases by 100% - (100% * .999) = 0.1% for every new message. Same as before, the larger the value of the poster homogeneity measure, the higher the risk of death, and the more messages, the lower the risk of death.

The regression coefficients for the first three levels of PFFirst2Weeks were relative to the reference category, which corresponded to the Constant production function type. The regression coefficients suggest that the hazard for channels with Ascending production functions is 0.692 times the hazard of channels characterized by Constant production functions, the hazard for channels with Descending production functions is 0.589 times the hazard of channels with Constant production functions, and the hazard for channels with Variable production functions is 0.623 times the hazard of channels with Constant production functions.

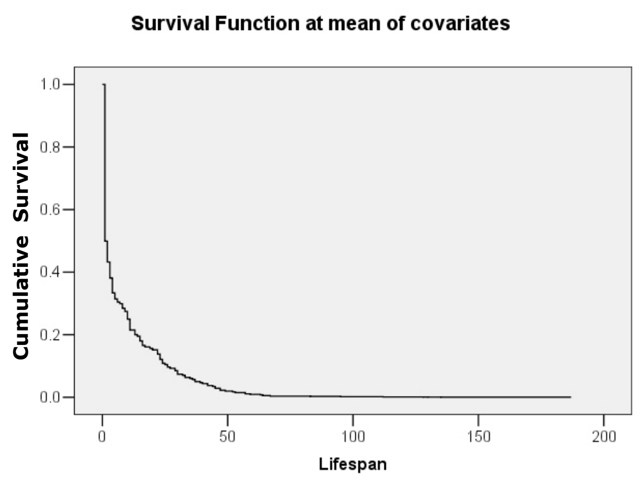

Figure 1. Survival Function for the First 2 Weeks of Life

The significance value of the regression coefficient for channels with Descending production functions was slightly larger than 0.10, so any observed differences between this category and the reference category could be due to chance. However, the other two regression coefficients had significance values smaller than 0.05, indicating that channels with Ascending and Variable productions functions were statistically different from channels with Constant production functions.

Following are charts of the production functions for the 2 week observation point. Similar charts with comparable patterns are available for the other time periods measured; however, they are not shown here. Lifespan is measured in days. Cumulative survival is a proportion.

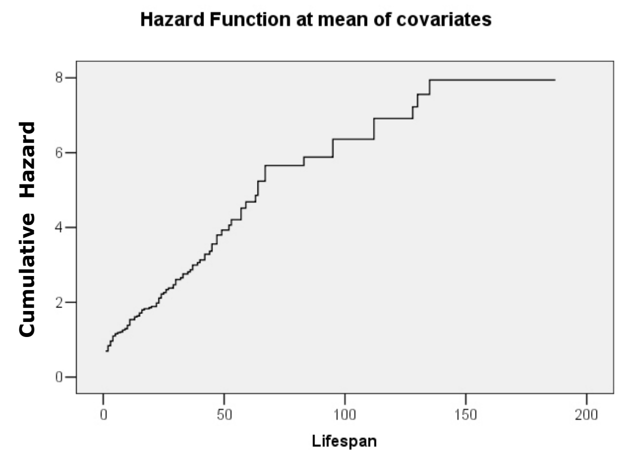

Figure 2. Hazard Function for the First 2 Weeks of Life

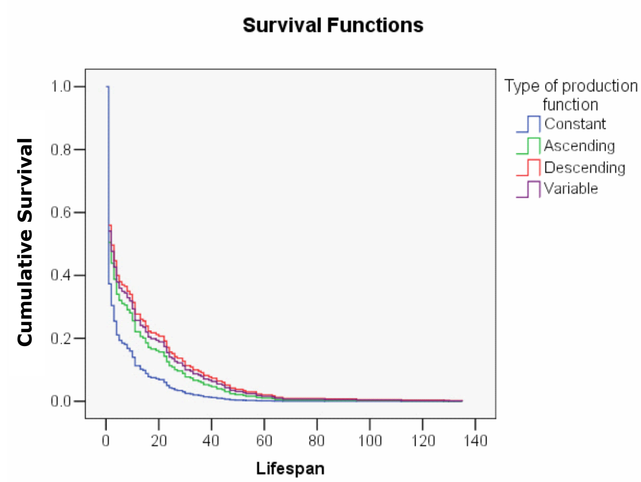

Figure 3. Survival Functions for Production Functions for the First 2 Weeks of Life

The basic survival curve shown in Figure 1 is a visual display of the model-predicted time to death for the "average" channel. The horizontal axis shows the time to event (the lifespan of the channel), while the vertical axis shows the probability of survival. Thus, any point on the survival curve shows the probability that the "average" channel will stay alive past that time. Past 65 days, the survival curve becomes less smooth because of the fewer channels that have survived for that long. The plot of the survival curves for each covariate pattern shown in Figure 3 gives a visual representation of the effect of the "production function type" category.

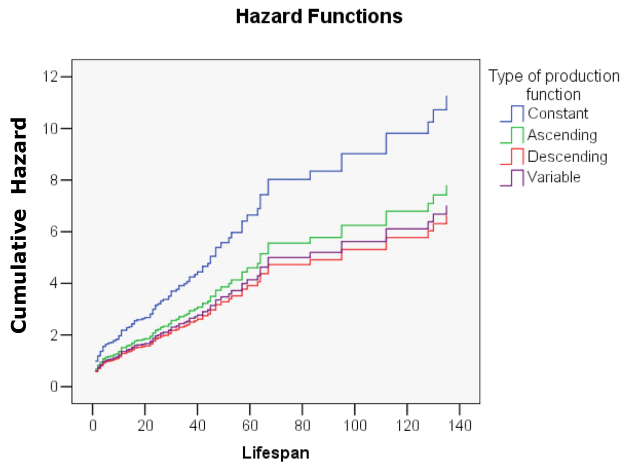

Figure 4. Hazard Functions for Production Functions for First 2 Weeks of Life

In this case the channels with Ascending and Variable production functions were statistically different from channels with Constant production functions. Channels with Variable production functions were the most likely to survive, followed by channels with Ascending production functions, while channels with Constant production functions were the least likely to survive.

The basic hazard curve shown in Figure 2 is a visual display of the cumulative model-predicted potential to die for the "average" channel. The vertical axis shows the cumulative hazard, equal to the negative log of the survival probability. Past 65 days, the hazard curve, like the survival curve, becomes less smooth, for the same reason. The plot of the hazard curves for each covariate pattern shown in Figure 4 gives a visual representation of the effect of the "production function type" category. Channels with Constant and Ascending production functions had higher hazard curves because they had a greater potential to not survive.

DISCUSSION

The present research investigated quantitatively the conditions at inception that best support an online community becoming and remaining successful over longer periods of time. We focused our empirical research on IRC chat channels. To predict the survivability of IRC channels, Cox regression models were created for four time intervals in the initial stages of the channels' lives: the first two hours, the first day, the first week, and the first two weeks. For each time interval 4 types of independent variables that characterized the channel activity were computed. These were: 1) user activity (messages, posters, etc.); 2) the trajectories of channel activity; 3) the heterogeneity of the channel's population; and 4) the channel's production functions as proposed suggested by Oliver and Marwell [15].

The number of messages was found to be a good indicator of the channels' likelihood of survival in three of the four cases. While the number of messages in the first two hours of life was not significant enough, the number of messages sent to a channel during its first day of life, first week of life and first two weeks of life were identified as good predictors by the Cox regression models. In all three cases, an increase of 1 message implied an increase of 0.1 percent in a channel's chance of survival, while 100 more messages sent to the channel implied an increase of 9.5 percent in the channel's chance of survival.

The correlation coefficients between the lifespan of channels and the number of users, posters, and messages were less than 0.5 during the first two hours and the first day, and grew above this value for the first week and the first two weeks. The numbers of users and of posters were not identified as statistically significant predictors in the regression. The implication is that mere quantity of participants is insufficient in describing a community and predicting its sustainability. This contrasts with prior research (e.g. [14]) which has focused simply on the number of people as a measure of critical mass and supports the premise offered by the *Critical Mass Theory*. It was this earlier finding that the number of participants on its own does not predict longevity that led to the notion that what constitutes critical mass is dependent on the social context [3].

The trajectories of channel activity did not predict the likelihood of channels' survival at all for any of the four intervals analyzed.

The homo/heterogeneity of the channels' user populations (measured by the PosterHG variable) was found to be a strong indicator for the channels' likelihood of survival by all four models. A channel was considered more homogeneous if its poster population stayed relatively constant as time passed, and more heterogeneous if its poster population changed significantly over time. Channels with heterogeneous populations were more likely to survive than channels with homogeneous populations. An increase of 1 percent in homogeneity in the first two hours of life implied a decrease of 1.3 percent in a channel's chance of survival. An increase of 1 percent in homogeneity in the first day of life implied a decrease of 1.5 percent in a channel's chance of survival. An increase of 1 percent in homogeneity in the first week of life implied a decrease of 2.6 percent in a channel's chance of survival. An increase of 1 percent in homogeneity in the first two weeks of life implied a decrease of 3.4 percent in a channel's chance of survival.

The Lifespan variable, which was the dependent variable, was negatively correlated with the PosterHG in all four intervals, and the correlation coefficients were quite high (Spearman's rho ranged between -.476 and -.658, $p<.01$). This shows that channels that survived longer were likely to be more heterogeneous than channels that survived for shorter periods. Although correlation does not imply causation, this relationship is worth exploring further.

These results clearly indicate that the longer a channel's population stayed homogeneous, the less likely that channel would survive. Lower heterogeneity may be an indication of lower poster turnover. Higher poster turnover may indicate persistent interest in the specific channel by more and more new posters. For example, a PosterHG value of 100 percent for a channel during any of the analyzed time intervals suggests that the channel had a constant/stable, homogeneous, group of posters throughout that interval. An explanation could be that a small group of people who resolved some chat among themselves and then did not meet again. The opposite of this case was when a channel had a lower value for the PosterHG variable. Such cases could be explained by more turnover of posters, therefore a persistent interest in the channel, hence the longer lifespan.

Thus, group heterogeneity, which is a social measure, emerges in the present study as the most prominent predictor of community success in the long term although it has not been identified previously as a critical success factor for community sustainability [6].

Critical Mass Theory defined production functions as the relationships between resources contributed by a group and the collective output of that group. In the case of IRC chat-channels, the number of users present in the channel was considered a surrogate measure for the group resources, while the number of messages was considered a surrogate measure for the amount of group success achieved. The shape of the channels' production functions was identified as a good predictor only for the longest analyzed interval. Four types of production functions were used: constant, ascending, descending, and variable. The results of Cox regression model for the first two weeks of life indicated that channels with constant production functions were the most likely to die, while channels with variable production functions were the most likely to survive. The likelihood of survival of channels with ascending production functions was slightly lower than that of channels with variable production functions. The model failed to produce statistically significant results for the channels with descending production functions, but this might have been due to the fact that there were only a few channels characterized by this type of production function, compared to the rest of the channels.

Overall, the Cox regression procedure produced a suitable model for predicting IRC channels' survivability. The use of separate blocks for fitting the model allowed guaranteeing that the production function categorical variable would be added to the final model, while still taking advantage of the stepwise techniques for choosing the other predictors. The best results were obtained when the regression model used the predictors that were computed for the immediate two-week period that followed a channel's creation.

The importance of the current findings is that they provide empirical evidence and support for Oliver and Marwell's *Critical Mass Theory*. Further, poster heterogeneity emerges as the central predictor of survival with the advantage that it is evident as early as within two hours of a channel's existence and continues to be the leading predictor in the longer term studied here. Activity in the form of messages sent is also a good predictor; however, it requires a longer time span in order to become reliable. Production functions require a much longer time to become significant predictors of survival. This situates the theories by Oliver and Marwell [15] and by Markus [12] within the actual practices of chat channels.

LIMITATIONS

The rigorous quantitative approach taken in this study does not provide answers when it comes to understanding users' motivations and intentions. For example, we can only assume why some channels were short lived but we did not interview their users to verify their perspectives. On the other hand, analyzing a vast amount of data allows us to observe a reliable representation of the actual IRC development.

CONCLUSION

The study presented here extends the existing literature by providing empirical data to support for the *Critical Mass Theory*'s application to online community systems. Further, the present study highlights the importance of group heterogeneity and group messaging activity as predictors of success rather than mere numbers of users. Future work may add the qualitative dimension by interviewing users.

REFERENCES

[1] Ackerman, M. S. and Palen, L. The Zephyr Help Instance: promoting ongoing activity in a CSCW system. In Anonymous *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground.* (). ACM New York, NY, USA, , 1996, 268-275.

[2] Andrews, D. C. Audience-specific online community design. Commun ACM, 45, 4 (2002), 64-68.

[3] Bradner, E., Kellogg, W. A. and Erickson, T. The adoption and use of BABBLE: A field study of chat in the workplace. In Anonymous *Proceedings of the Sixth European conference on Computer supported cooperative work.* (). Springer, , 1999, 139-158.

[4] Hardin, R. *Collective action.* Johns Hopkins University Press, , 1982.

[5] Hiltz, S. R. and Turoff, M. Structuring computer-mediated communication systems to avoid information overload. (1985).

[6] Iriberri, A. and Leroy, G. A life-cycle perspective on online community success. (2009).

[7] Jones, Q. Applying cyber-archaeology. In Anonymous *ECSCW 2003: Proceedings of the Eighth European Conference on Computer Supported Cooperative Work, 14-18 September 2003, Helsinki, Finland.* (). Kluwer Academic Publishers, , 2003, 41.

[8] Jones, Q. and Rafaeli, S. Time to split, virtually:'Discourse Architecture'and 'Community Building'as means to creating vibrant virtual publics. Electronic Markets: The International Journal of Electronic Commerce and Business Media, 10, 4 (2000), 214-223.

[9] Jones, G. Q. and Rafaeli, S. User population and user contributions to virtual publics: A systems model. In Anonymous *the ACM International Conference on Supporting Group Work (Group99).* (Phoenix, AR, USA,). ACM Press, , 1999, 239-248.

[10] Kling, R. and Courtright, C. Group behavior and learning in electronic forums: A sociotechnical approach. The Information Society, 19, 3 (2003), 221-235.

[11] Licklider, J. C. R. and Taylor, R. W. The computer as a communication device. Science and technology, 76, 21 (1968), 621-626.

[12] Markus, M. L. Toward a" critical mass" theory of interactive media: Universal access, interdependence and diffusion. Communication Research, 14, 5 (1987), 491.

[13] Moldovan, M., Raban, D. R., Butler, B. S. and Jones, Q. G. Empirical Evidence of Information Overload Constraining Chat Channel Community Interactions. In Anonymous *ACM conference on Computer Supported Cooperative Work.* (San Diego, CA,). , 2008, 323-332.

[14] Morris, M. and Ogan, C. The Internet as mass medium. Journal of Computer-Mediated Communication, JCMC, 1, 4 (1996), 9/23/2009.

[15] Oliver, P., Marwell, G. and Teixeira, R. A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action. American Journal of Sociology, (1985), 522-556.

[16] Olson, M. *The logic of collective action: Public goods and the theory of groups.* Harvard Univ Pr, , 1971.

[17] Palme, J. *Electronic mail.* Artech House Boston, , 1995.

[18] Preece, J. Supporting Community and Building Social Capital-Introduction. Communications of the ACM-Association for Computing Machinery-CACM, 45, 4 (2002), 36-39.

[19] Raban, D. R. and Rafaeli, S. Investigating Ownership and the Willingness to Share Information Online. Computers in Human Behavior, 23(2007), 2367-2382.

[20] Rafaeli, S. The electronic bulletin board: A computer-driven mass medium. Soc. Sci. Comput. Rev., 2, 3 (1984), 123.

[21] Schelling, T. C. *Micromotives and macrobehavior.* WW Norton, New York, 1978.

[22] Schoberth, T., Preece, J. and Heinzl, A. Online Communities: A Longitudinal Analysis of Communication Activities. In Anonymous *HICSS'36.* (). , 2003.

[23] Whittaker, S. Talking to strangers: An evaluation of the factors affecting electronic collaboration. In Anonymous *Proceedings of the 1996 ACM conference on Computer supported cooperative work.* (). ACM New York, NY, USA, , 1996, 409-418.