The Effect of User Features on Churn in Social Networks

Marcel Karnstedt[1], Matthew Rowe[2], Jeffrey Chan[1], Harith Alani[2], Conor Hayes[1]

[1]Digital Enterprise Research Institute (DERI)
NUI Galway, Ireland
{firstname.lastname}@deri.org

[2]Knowledge Media Institute (KMI)
The Open University, Milton Keynes, UK
{m.c.rowe, h.alani}@open.ac.uk

ABSTRACT
Social sites and services rely on the continuing activity, good will and behaviour of the contributors to remain viable. There has been little empirical study of the mechanisms by which social sites maintain a viable user base. Such studies would provide a scientific understanding of the patterns that lead to user churn (i.e. users leaving the community) and the community dynamics that are associated with reduction of community members – primary threats to the sustainability of any service. In this paper, we explore the relation between a user's value within a community - constituted from various user features - and the probability of a user churning.

Categories and Subject Descriptors
H.4.3 [**Information Systems Applications**]: Communications Applications—*Bulletin boards*; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms
Experimentation, Human Factors

1. CHURN IN SOCIAL NETWORKS
In studies of churn behaviour of customers of telecommunication networks, a user's probability of churning has been linked to the churning behaviour of neighbours in his/her social network. This has recently also been observed in online social networks [5].

In this paper, we examine relationships between user value and churn. By churn we refer to the loss of users, as one indicator for decreasing community value, implicitly encoding the idea that a user no longer finds a service useful or valuable and has moved elsewhere. Building on our previous work [5], we explore the correlation between behavioural and structural user features that are commonly used to describe user value and churn probability and influence, identifying key indicators of churn within a community. For our experiments over a year's worth of data, we profile contributors in an online bulletin board by extracting salient behavioural and structural features. Our approach employs time-series analysis, identifying links between certain user value features and their evolution with time, and the probability of an individual leaving the community. Our hypothesis is that users which display different behavioural, content and structural characteristics in the underlying social network will tend to have different influence on churn. By this we identify features of contributors that are implicitly recognised by other users as contributing to the value of the community. This provides an important contribution to the analysis of the relationship between user value, user churn and community value in general. It produces an understanding of the behavioural patterns associated with the loss of community members, eventually enabling community hosts to identify, early-on, that users may leave the community.

We base our empirical analysis on data from the popular Irish forum site *boards.ie*[1] over the course of the year 2006. The same data was used in [5] to assess the used churn definition and inspect different window sizes and threshold values used in the definition. As we describe below, for the analysis we also have to inspect parts of the data from 2005. The reply structure, which we use to define communication relations between users, is explicitly available in the data.

We have structured the paper as follows: Section 2 lists related work within the field of churn analysis and how user value is defined within the literature. Section 3 describes our feature engineering process and the features that we use for analysis that define user value within an online community. Section 4 describes our experiments and the three analysis that we conducted: 1) correlating global churn with user value; 2) correlating forum-dependent churn within user value within sub-communities, and; 3) correlating neighbourhood churn within sub-communities with an individual's likelihood to churn. Section 5 discusses the findings from our analysis and Section 6 finishes the paper with the conclusions that we were drawn from this work.

2. RELATED WORK
To date, churn has been mostly analysed and discussed in the context of monetary services, most prominently in the telecommunication sector, where churn is commonly understood as a total defect of a customer. However, there are also works inspecting partial defects [1], which we see as a

WebSci '11, June 14-17, 2011, Koblenz, Germany.

[1]http://boards.ie

crucial characteristic for churn in social networks. While most works like [1] base their analysis and approaches for churn prediction on a classical feature-based approach [8], several recent works already take up the idea of social network analysis in this context. [3] models churn as a spread of influence. A similar approach is taken in [6] to model churn for multi-player online games. Thus, this is one of the first works discussing churn in a domain close to online social networks.

In this work, we analyse the correlation between churn probability and a set of features that are commonly used to describe user and community value, trying to explain motivations of users to contribute to and to stay in a community. The literature citing such definitions and quantitative analysis in the setting of online communities is limited. Several motivations for contribution to digital social networks have been proposed [7], where a key observation of user behaviour in online networks is that users, with the exception of spammers, make contributions to online discourse without expecting any immediate return. An article by Clay Shirky in [10] describes how communities function through '*intercasts*', where information is shared and content is unique to the community. The findings parallel the design of a discussion board as analysed in our work. [9] lists a range of intuitive measures that make up the value of an individual in an organisation, which is related to the value of users in a community. Our notion of community churn relates to turnover and retention as used in [9] – indicating that health signifiers of a community and organisation correlate with user retention. Besides such rather behavioural measures, assessing the structural features of networks provides a useful technique for gauging user value through numerical values derived from social network analysis. Work described in [4] cites the utility of such metrics when measuring the value of users within consumer communities.

In this work, we apply ideas and techniques from all the above mentioned and similar works. [5] is one of the first works inspecting the notion of churn in social networks. It presents an exhaustive discussion and a comprehensive list of related work regarding different notions of churn, reasons for churn, approaches to predict it, and much more. The proposed definition of churn and first insights into factors influencing churn build the basis of the work in hand.

3. FEATURE ENGINEERING

We formalise our analysis as an assessment of the correlation between a given user (v_i) and their churn probability and features at a given point in time: t_k. To determine the user churn probability we use the activity-based definition from [5]. It is based on comparing user activity from two time windows: a previous activity window and a churn window. Based on our initial analysis of the boards.ie dataset, we found that setting both the previous activity window and the churn window to 13 weeks identified churners in the most pronounced way and reduced noise – this is described in [5]. As our goal is to compare the probability of individual churn with the same user's features, we also require a feature window to be set. The feature window is the window of analysis from which we draw the past posts by a given user up until a given point in time (t_k). We decided on using a longer window for analysis than the previous activity window to capture a broader spectrum of data from which user features could be compiled. We set the length of this window to be 26 weeks, thus covering the same length of time span as the previous activity window and churn window combined – but stopping prior to the churn window. Figure 1 summarises our window settings. At a given point in time t_k we want to measure the correlations between a given user's churn probability and her features. The previous activity window is defined as $(t_k - n) \rightarrow (t_k - 1)$, the churn window as $t_k \rightarrow (t_k + m - 1)$. The feature window is composed of 26 weeks prior to t_k: $(t_k - (n + m)) \rightarrow (t_k - 1)$ - setting both n and m to 13 weeks.

For our analysis we assess user churn probability and features throughout the year 2006, starting on 1st January 2006 and calculating the churn probability and features at weekly increments. Therefore, to calculate the user features for the first time stamp, we require data from the second half of 2005 for our feature window and the last quarter of 2005 for our previous activity window. It is worth noting also that we do not calculate the churn probability of every week in the year, we only run this analysis up until week 39, given that this is the point in time where the churn window reaches the end of the year 2006.

3.1 Dependent Variable: Churn Probability

[5] defines churn as a binary assessment of a user's activity in the previous activity window and churn window, stating that a user had *churned* should their activity drop below a given rate:

$$\mu_C(v_i) \leq T(S).\mu_{PA}(v_i) \quad (1)$$

$\mu_C(v_i)$ denotes the average activity in the churn window (C), $\mu_{PA}(v_i)$ denotes the average activity in the previous activity window (PA) and $T(S)$ defines a system-specific parameter in the range $0 \leq T(S) \leq 1$. We can rewrite this as a probability estimate as follows:

$$P(churn|v_i) = \begin{cases} 0 & \mu_C(v_i) \geq \mu_{PA}(v_i) \\ 1 - (\frac{\mu_C(v_i)}{\mu_{PA}(v_i)}) & otherwise \end{cases} \quad (2)$$

The above equation returns a probability of 0 if the activity of the user in the churn window is the same or greater than the activity in the activity window. Otherwise, it derives the proportion of activity in the churn window with regards to the activity window and converts this to a churn probability. The system specific parameter ($T(S)$) now becomes a threshold against which we can compare the churn probability, and should it be exceeded, declare v_i as having churned.

3.2 Independent Variables: User Features

In absence of explicit social connections (e.g, '*following*' or '*friending*'), we represent the communication interaction between users as a weighted, directed graph $G(V, E)$, the *reply graph*. This is denoted by $G(V, E)$, where V is the set of vertices and E is the set of edges between a pair of vertices. Each vertex $v \in V$ represent a user in a forum, and a directed edge $e(i, j) \in E$ exists from user v_i to user v_j if user v_i has replied to a post of user v_j in a thread in the forum. We associate the number of posts between two users as the edge weight. In the following, we define the features used in this work and briefly state the intuition behind them.

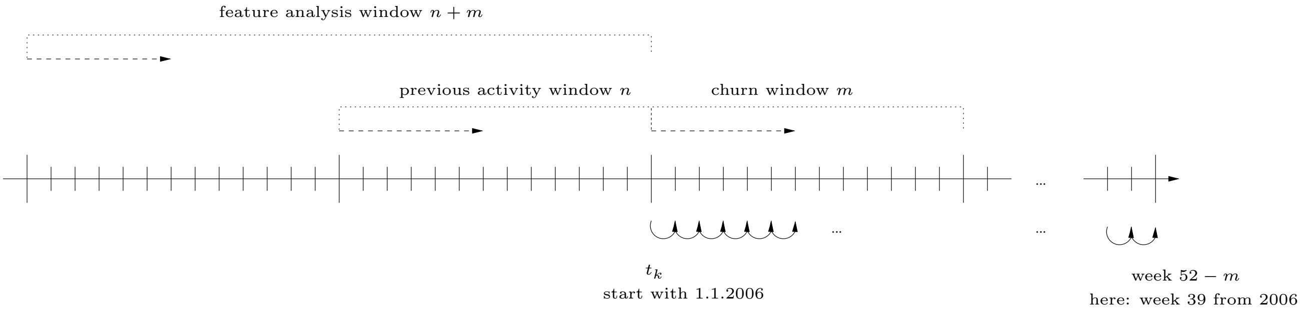


Figure 1: Illustration of windows for feature extraction and analysis

In-degree. In-degree measures the number of incoming connections to a given user v_i. To measure the in-degree of a given user (v_i) we count the number of unique users that have replied to that user in the past-6 months.

Out-degree. Out-degree measures the number of outgoing connections from a given user v_i. We derive this measure in a similar manner to a user's in-degree, by counting the number of unique users that the user v_i has replied to over the past 6-months.

Closeness Centrality. Closeness centrality measures the importance of a user based on their location in the reply graph. A central user will tend to have high closeness centrality; i.e. if the reply graph was thought of a information passing network, then rumours initiated by a central user will spread to the whole network quicker. Let $d_{i,j}$ be the length of the shortest path between vertices v_i and v_j. Then average distance between vertex v_i and all vertices is given by:

$$l_i = \frac{1}{|V|}\sum_{j \in V} d_{i,j}$$

The closeness centrality is defined as the inverse of l_i.

$$C_i = \frac{1}{l_i} = \frac{|V|}{\sum_{j \in V} d_{i,j}}$$

Betweenness Centrality. Betweenness centrality is another importance measure. Users with high betweenness tend to be conduits or brokers between communities. Let $\gamma_{x,y}$ be the number of shortest path between vertices v_x and v_y. Let $\gamma_{x,y,i}$ be the number of those paths where v_i lies on the path, and $v_i \neq v_x$ and $v_i \neq v_y$. Then the betweenness for vertex v_i is defined as:

$$B_i = \sum_{x,y \in V} \frac{\gamma_{x,y,i}}{\gamma_{x,y}}$$

For brevity, hereafter we shall refer to closeness centrality as centrality and betweenness centrality as betweenness.

Reciprocity. Reciprocity measures the average time it takes for a post of a user to be replied to. It provides an indication of how important a user is, and the type of posts the user posts; the assumption being that different types of posts have different response time. Let pst_x denote a post, $Pst(t_i, t_2)$ denote all posts written over the period $[t_1, t_2]$ and $Pst_i(t_1, t_2)$ denote the set of all posts written by user v_i over the period $[t_1, t_2]$. Let $r(pst_x, pst_y, t_{xy})$ denote that pst_y is a reply of pst_x and there was a delay of t_{xy} time units (we use minutes) between the posting times of the posts. Then the reciprocity for user v_i is defined as:

$$rep_i(t_1, t_2) = \frac{1}{R}\sum_{r(pst_x, pst_y, t_{xy})} t_{xy}$$

where $R = |r(pst_x, pst_y, t_{xy})|$, $pst_x \in Pst_i(t_1, t_2)$, and $pst_y \in Pst(t_1, t_2)$.

Average post in initiations. This measures the average length of discussion/conversation that occurs on threads initialised by a user. Let thr_l denote a thread, and $|thr_l|$ the number of posts in thread. Let $init_i(t_1, t_2)$ denote the set of threads initialised by user v_i over period $[t_1, t_2]$. Then the average length of threads initialised by user v_i is

$$\frac{1}{|init_i(t_1, t_2)|}\sum_{thr_l \in init_i(t_1, t_2)} |thr_l|$$

Average post in participations. This measures the average length of discussion/conversation that occurs on threads that a user participates in. It gives an indication to what types of threads a user typically posts in; short threads could be Q&A type of threads while longer threads could be discussion type of threads. Let $part_i(t_1, t_2)$ denote the set of threads participated by user v_i over period $[t_1, t_2]$. Then the average length of threads participated by user v_i is

$$\frac{1}{|part_i(t_1, t_2)|}\sum_{thr_l \in part_i(t_1, t_2)} |thr_l|$$

Popularity. Popularity measures the percentage of posts written by a user that are replied to. A user who is more popular will be more likely to get a reply to a post they made. The popularity for user v_i is defined as

$$pop_i(t_1, t_2) = \frac{R}{|Pst_i(t_i, t_2)|}$$

where $R = |r(pst_x, pst_y, t_{xy})|$, $pst_x \in Pst_i(t_1, t_2)$, and $pst_y \in Pst(t_1, t_2)$.

Initialisation. Initialisation measures the popularity of threads initialised by a user. The more popular a user, the more likely a thread they initialised will receive at least one reply. Then initialisation of a user v_i is defined as

$$init - pop_i(t_1, t_2) = \frac{|\{thr_l | thr_l \in init_i(t_1, t_2) \wedge |thr_l| > 1\}|}{|init_i(t_1, t_2)|}$$

Polarity. Polarity captures the average sentiment – i.e. positive/negative feeling – of a user's posts over the past 6 months. We gather the collection of posts that a user has made – i.e. $p \in P_{v_i}$ – and measure the polarity of each post using Sentiwordnet's sentiment lexicon.[2] We then take the average of the polarity measure of each post in the collection. To measure the polarity of a single post we use the following formula, where c is the number of unique terms in post p, the function $pos(T)$ returns the positive weight of the term T from the lexicon and $neg(T)$ returns the negative weight of the term:

$$polarity(p) = \frac{1}{c}\sum_{k=1}^{c} pos(T) - neg(T)$$

4. EXPERIMENTS

In this section, we describe three analysis tasks and the experiments that accompanied each. The first analysis seeks to correlate a user's value with the likelihood that the user will churn in a global setting – i.e. analysing the entire boards.ie platform for one year. The second task analyses forum specific behaviour, correlating user features with churn probability in four different forums. And finally, the third task explores the neighbourhood effects on churn in each of these four forums and the differences that community dynamics have upon such correlations.

4.1 Global Churn Analysis

For the global churn analysis, we use the previously described features, but omit centrality and betweenness due to their computational complexity and run time. The data from the inspected time frame spans 486 forums. 32.826 users show activity over all these forums, summing up to a total of 2.363.404 posts, where 2.168.546 of them are replies to another post. Our initial experiments involved regression models induced from the data, seeking to achieve a high Coefficient of Determination (R^2) from such a model to explain correlations. After repeated analysis with various different regression models, testing Linear Regression, Least Median of Squares, Isotonic Regression and Support Vector Regression, we were unable to achieve satisfactory R^2 levels, thus making conclusions drawn from correlations in the data hard to justify – e.g. we achieved a low R^2 value of 0.0069 using Linear Regression.

Due to the limitations of regression analysis in this instance, we altered the correlation task to one supported in a binary classification setting. We replicated the binary churn decision from [5] by creating a dataset for all users who participated on the site during 2006, assigning each user at a given week a binary class label designating them as either *pos* (churner) or *neg* (non-churner) We varied the threshold σ that the binary decision is based on between three values of $\sigma = \{0.2, 0.5, 0.7\}$.

Based on these three thresholds, we created three different datasets and divided each dataset into a training and testing set using an 80/20% split. The J48 decision tree classifier was then trained on the former split and tested on the latter. We assessed the classification performance using the standard measures of precision, recall and f-measure (F_1) – setting $\beta = 1$ for an equal weighting of precision and recall – and reported on the κ coefficient for agreement between predictions and the class labels. Once we had identified the threshold that produced the highest F_1 value, we then took the training split from that threshold setting and analysed the correlation of features with class labels using Information Gain Ratio to calculate the *worth* of each feature and boxplots to visualise the feature distribution with regards to churners and non-churners.

[2] http://sentiwordnet.isti.cnr.it/

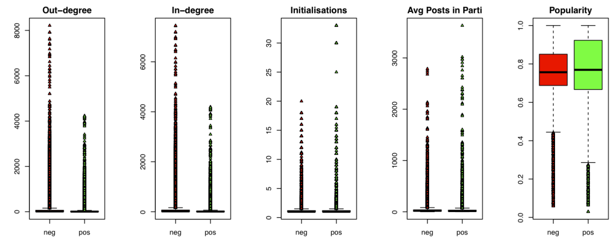


Figure 2: Boxplots of the correlation between the top-5 features on the Boards.ie platform, ranked by Information Gain Ratios, with respect to Churner (pos) and Non-churner (neg) class labels

Results. Table 1 shows the classification results from the churner identification at various thresholds. The results show that as the threshold increases we see an improvement in all evaluation measures, indicating that for lower values of σ additional false positives and false negatives are produced. The kappa statistic (κ) also increases as the threshold increases, indicating a greater agreement between the predictor's decisions and the labels in the data.

Table 1: Results from Churner prediction for different churn thresholds. Note that P denotes precision and R denotes recall.

Threshold	P	R	F_1	κ
0.2	0.638	0.639	0.635	0.266
0.5	0.668	0.666	0.649	0.286
0.7	0.734	0.741	0.733	0.410

In order to assess the contribution of each feature in terms of discriminating between churners and non-churners, we took the training split of the dataset compiled using the best performing threshold in terms of F_1 levels – $\sigma = 0.7$. Using this dataset we computed the Information Gain Ratio (IGR) of each feature with respect to differentiating between churners and non-churners. Table 2 shows the results from this analysis with features ranked by their IGR. The out-degree and in-degree of users top the list, indicating that key differences in these features can be used to segment churners from non-churners. Initialisations comes third, indicating that the number of posts started by a given user is also a good indicator of separating churners from non-churners.

Although the IGR ranking provides an insight into important features when identifying churners, the ranges of the features are not explained. To do this we analyse the boxplots of features in the training split with respect to the two class labels. Figure 2 shows these plots, where for in-degree and out-degree we observe higher values as being correlated with non-churners. The differences between churners and non-churners for Initialisations and Average Posts in Participations is not as evident. However, for popularity there is a clear difference in the distributions between the churner and non-churner class labels, where the popularity for non-

Table 2: Features ranked by Information Gain Ratio wrt Churner class label when analysing global churn. The feature names is paired with its Information Gain Ratio in brackets.

Rank	Feature
1	Out-degree (0.0076)
2	In-degree (0.0072)
3	Initialisations (0.0056)
4	Avg Posts in Parti (0.0042)
5	Popularity (0.0042)
6	Polarity (0.0028)
7	Reciprocity (0.0024)
8	Avg Posts in Initi (0.0021)

churners is less distributed than for churners. This follows intuition given that a consistent level of popularity is associated with users who stay and participate within the community, while those who leave a community will have more sporadic behaviour.

4.2 Per Forum Churn

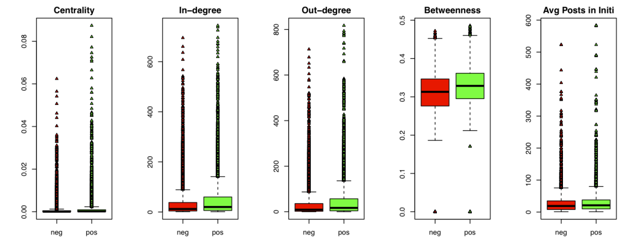
In the per forum analysis, we seek to identify common behavioural patterns across different forums and idiosyncrasies unique to specific forums. We use the same features as in the global churn analysis, but include centrality and betweenness in our analysis this time. To provide a variety of forums to analyse we chose the following four as follows:

- *Highest Activity:* forum 7 (After hours - general discussion forum): 6.159 users, 232.880 posts, 226.207 replies
- *Lowest Activity:* forum 224 (Nihongo - Japanese language forum): 60 users, 122 posts, 96 replies
- *Median Activity:* forum 512 (Prime Time Cartoons - television discussion forum): 459 users, 1.909 posts, 1.758 replies
- *Mean Activity:* forum 524 (World of Warcraft - computer game discussion forum): 411 users, 5.787 posts, 5.345 replies

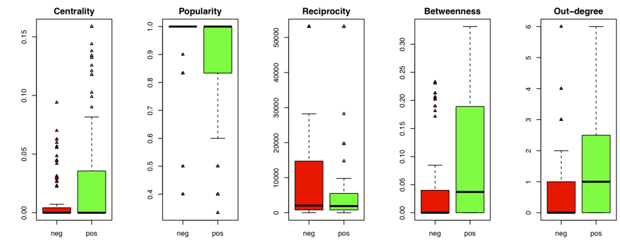
Table 3: Results from Churner prediction in different forums where the churn class label is assigned based on the churn probability exceeding the stated threshold

Forum	Threshold	P	R	F_1	κ
7	0.2	0.724	0.727	0.722	0.422
	0.5	0.658	0.657	0.656	0.312
	0.7	0.772	0.773	0.772	0.525
224	0.2	0.753	0.754	0.749	0.482
	0.5	0.724	0.725	0.723	0.439
	0.7	0.724	0.725	0.723	0.439
512	0.2	0.791	0.790	0.790	0.559
	0.5	0.787	0.787	0.787	0.572
	0.7	0.756	0.755	0.755	0.510
524	0.2	0.755	0.743	0.745	0.481
	0.5	0.695	0.539	0.425	0.079
	0.7	0.773	0.771	0.766	0.521

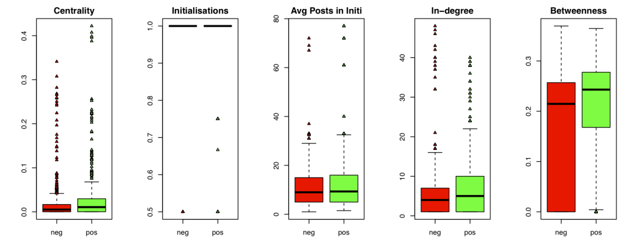
Similar to the global churn analysis, we observed low R^2 values when regressing the user features against an individual's churn probability – i.e. yielding 0.0116, 0.146, 0.0132 and 0.0204 for forums 7, 225, 512 and 514 respectively. Therefore, we repeated the binary classification task as described in the previous section but on a per-forum basis, building three datasets for each of the four forums. Again, users were labelled as either *pos* (churner) or *neg* (non-churner)



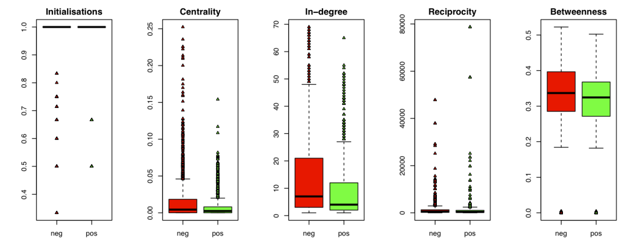
(a) Forum 7



(b) Forum 224



(c) Forum 512



(d) Forum 524

Figure 3: Boxplots of the correlation between the top-5 features in each forum, ranked by Information Gain Ratios, with respect to Churner (pos) and Non-churner (neg) class labels

depending on a given threshold of $\sigma = \{0.2, 0.5, 0.7\}$. We split each dataset using an 80/20% split and trained the J48 decision tree classifier on the former split and tested it on the latter. The training split from each forum was selected for the threshold that achieved the maximum F_1 value, and analysed the correlations of features within this split with class labels.

Results. The results from identifying churners in the various forums are shown in Table 3. We observe varying results for the different forums when σ is altered. For example, in the high activity forum (7) for general discussions, $\sigma = 0.7$ maximises f-measure, while this value is $\sigma = 0.2$ for the low activity forum (224). We also observe that in forum 7 where $\sigma = 0.5$ we see a reduction in performance, indicating that

Table 4: Features ranked by Information Gain Ratio wrt Churner class label. The feature name is paired within its Information Gain Ratio values in brackets.

Rank	Forum 7	Forum 224	Forum 512	Forum 524
1	Centrality (0.0065)	Centrality (0.1574)	Centrality (0.1356)	Initialisations (0.0369)
2	In-degree (0.0062)	Popularity (0.1310)	Initiations (0.0523)	Centrality (0.0274)
3	Out-degree (0.0059)	Reciprocity (0.1029)	Avg Posts in Initi (0.042)	In-degree (0.0265)
4	Betweenness (0.0058)	Betweenness (0.0978)	In-degree (0.0201)	Reciprocity (0.0233)
5	Avg Posts in Parti (0.0029)	Out-degree (0.0818)	Betweenness (0.0198)	Betweenness (0.0188)
6	Polarity (0.0029)	In-degree (0)	Reciprocity (0.019)	Popularity (0.0174)
7	Popularity (0.0029)	Initialisations (0)	Popularity (0)	Polarity (0.0150)
8	Avg Posts in Initi (0.0028)	Polarity (0)	Out-degree (0)	Out-degree (0.0146)
9	Reciprocity (0.0024)	Avg Posts in Initi (0)	Polarity (0)	Avg Posts in Parti (0)
10	Initialisations (0)	Avg Posts in Parti (0)	Avg Posts in Parti (0)	Avg Posts in Initi (0)

differentiating between churners and non-churners is difficult at this threshold. For the the median activity forum (512) $\sigma = 0.2$ maximises f-measure and in the mean activity forum (524) $\sigma = 0.7$ maximises f-measure. In a similar manner to forum 7, the mean activity forum also yields a significant reduction in f-measure for the threshold setting of $\sigma = 0.5$, indicating, once again, the problem in predicting churners in this *grey area*.

Our results in Table 3 show the varying performance of identifying churners at differing churn thresholds. Taking the value of σ for each dataset that achieved the highest F_1 value for each forum, we then analysed the contribution of each feature using Information Gain Ratio (IGR) - as shown in Table 4 - and compiled boxplots shown in Figure 3 for the top-5 ranked features for each forum. We note that for centrality in three of the forums higher values correlate with the churner class label, while lower values are associated with non-churners. Centrality measures the information flow through a user within the community's interaction graph. Therefore, in three of the forums, the greater the information flow through the individual the more likely they are to churn for each of the specific thresholds. The exception to this is the mean activity forum 524.

For several other features we also observed idiosyncratic behaviour in the mean activity forum that was inconsistent with the other three analysed forums. For instance, for betweenness we find a correlation between higher values and churners, while for the mean activity forum, higher betweenness is correlated with non-churners – again emphasising the difference in community dynamics between the forums. For in-degree we also note that for forum 524 higher values are associated with non-churners, while for the other forums the opposite is the case.

4.3 Per Forum Neighbourhood Churn

For our third and final analysis task, we wished to explore the effects of a user's neighbourhood on his/her churn probability. Our intuition behind this analysis was that as more of a given user's neighbours churn, then the likelihood that the user will churn will also increase. We repeated the analysis over the previously described 4 discussion forums to investigate forums with a range of activity levels. To explore the correlation of neighbourhood churn with individual churn we explored two different types of neighbourhood churn probability: *unweighted* and *weighted*.

Unweighted Neighbourhood Churn. To derive the unweighted neighbourhood churn probability for a given user in a given forum, we begin by constructing a reciprocal network for v_i. This is derived by analysing the reply-to network from the previous 6 months of interactions within the given forum. We build an edge between v_i and another user v_j if the two users have both replied to one another. Following compilation of the egocentric unweighted reciprocal network for v_i, we then derive the average churn probability for users in the network as follows: for each user that v_i is connected to, we look up the churn probability of that user in the forum for the same week. Note that we do not account for lagged aspects in neighbour churn here due to the weekly time increments. If we had recorded churn probabilities over day increments, then lagged behaviour would be an important factor. We then take the average for all neighbours and record this as the *unweighted neighbourhood churn probability.*

Weighted Neighbourhood Churn. To derive the weighed neighbourhood churn probability for a given user in a given forum, we repeat the same process as above by constructing the egocentric network for the user. However, in this instance we weight edges between users based on interaction frequency. To do this we count the number of messages sent from v_i to each network member v_j and the number of messages received from each network member v_j. Then, we half the sum of these counts. This provides the weight w_j between user v_i and v_j born of interaction frequency. To provide a normalised weighting scheme for the reciprocal network, we convert the network weights ($w_j \in W$) into an influence distribution. An influence parameter λ_j describes the influence that user v_j has in the egocentric reciprocal network, where all influence factors in this network sum to 1 – i.e. $\sum_{j=1}^{\Lambda} \lambda_j = 1$. To convert each weight between v_i and v_j into an influence factor, we use the following normalisation scheme: $\lambda_j = w_j / \sum_{k=1}^{W} w_k$. From our influence factors we then compile the weighted neighbourhood churn probability as above, but in this instance we weight each neighbour's churn probability using their influence factor λ_j:

$$P(weightednchurn|v_i) = \frac{1}{|\Lambda|} \sum_{j=1}^{|\Lambda|} \lambda_j . P(churn|v_j) \qquad (3)$$

For our experiments, we repeated the same process as in the previous section, using individual datasets for each forum. However, in this task we utilised Linear Regression models. We set the dependent variable as the churn probability in the forum and set two independent variables: *unweighted neighbourhood churn probability* and *weighted neighbourhood churn probability.* Through the use of Linear Regression models, our analysis would then identify dependencies and correlations in the data, the significance of which would be empirically measured using the coefficient of determination (R^2) of the model.

Results. The results of this analysis are shown in Table 5 and Table 6 for forums 224 (least activity) and 512 (median activity), respectively. We only report on the results from these two forums due to their R^2 levels indicating a correlation in the data between the dependent variable and the predictors.

Table 5: Model results Forum 224

Feature	Coefficient	St' Error	t-Value	P(x >t)
Weighted	-0.345	0.076	-4.543	6.56E^{-06} ***
Unweighted	1.047	0.059	17.895	<2E^{-16} ***
Res. St Err: 0.3581 Mult R^2: 0.4319 $F_{2,685}$: 260.4 p-value: $< 2E^{-16}$				
Signif. codes: p-value < 0.001 *** 0.01 ** 0.05 * 0.1 . 1				

For forum 224 (least activity), Table 5 demonstrates a fair correlation in the model – defined by the multiple R^2 level of 0.4319. We observe that as an individual's churn probability increases, in the model, we see a reduction in the weighted neighbourhood churn probability, while the unweighted churn probability of neighbours increases. For the latter feature the t-test indicates the relationship as being of greater significance – characterised by the lower p-value than the weighted churn probability. This indicates that as a user's neighbours churn, they are more likely to churn also, given that their neighbours' activity drops influencing their activity to reduce also. For the weighted churn, the influence factors appears to have a lesser effect on churn probability. One would expect that as more important nodes in the reciprocal network churn, then this would impact upon the user's churn probability and lead to a reduction in activity

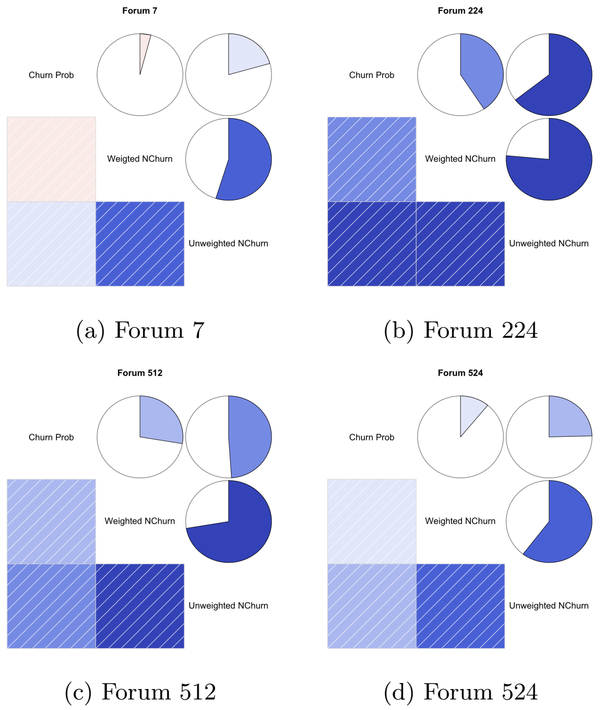
For forum 512 (median activity) we observe a lower R^2 value of 0.2519, see Table 6. The coefficients in the model, however, exhibit similar behaviour to Forum 224. As an individual's churn probability increases, the weighted neighbourhood churn probability decreases in the model, while the unweighted churn probability increases.

Table 6: Model results Forum 512

Feature	Coefficient	St' Error	t-Value	P(x >t)
Weighted	-0.440	0.036	-12.148	<2e-16 ***
Unweighted	0.928	0.021	45.249	<2E^{-16} ***
Res. St Err: 0.4151 Mult R^2: 0.2519 $F_{2,8720}$: 1468 p-value: $< 2E^{-16}$				
Signif. codes: p-value < 0.001 *** 0.01 ** 0.05 * 0.1 . 1				

To this point, we have considered the correlations of variables within the same linear model and the predictive power that the model has on an individual's churn probability. To explore the correlation of an individual's churn probability with each of the neighbourhood churn probabilities, we examined the correlograms produced from each forum's data, shown in Figure 4. Correlograms allow the correlation between pairs of variables to be visualised with pie charts on the upper right side and line diagrams on the lower left – the former denoting the strength of the correlation and the latter denoting the direction (and sign) of the correlation (i.e. as one variable increases, so too does the compared variable).

Taking forum 7, as shown in Figure 4(a), we see that a marginal correlation exists between the churn probability and the weighted neighbourhood churn probability, while a



(a) Forum 7 (b) Forum 224

(c) Forum 512 (d) Forum 524

Figure 4: Correlograms of the correlation of neighbourhood churn - both weighted and unweighted - with individual churn probability

slightly larger correlation exists between the churn probability and the unweighted neighbourhood churn probability. Analysing the correlogram for forum 524, shown in Figure 4(d), we see a similar effect where the correlation between the individual churn probability is only marginally correlated with the neighbourhood churn probability. In both these forums the activity rates are higher than the other two analysed forums. This, in turn, has an effect on the churn probability of the user and how it is effected by his/her neighbours. If we consider that forums with less frequent interactions are driven by a stronger community feel, where the absence of one community member is noticed by others, then the effect of neighbourhood churn will be greater. Conversely, in forums with greater discussion activity users interact with many different users, thereby making the churn of one user have an insignificant effect on the activity in the community, and therefore the likelihood of individual churn.

Focussing on the two remaining forums 224 and 512, we observe much greater correlations between individual churn probability and the neighbourhood churn probabilities. For forum 224, as shown in Figure 4(b), we note a significant correlation between the churn probability and the unweighted churn probability of neighbours. We also observe the same for forum 512. As we stated above, lower activity forums exhibit a stronger relation between the likelihood that a user churns when his/her neighbours leave the network.

5. DISCUSSION

There are of course numerous internal and external parameters that can influence churn in online communities. In this work, we focused on a set of user and content features that are pertinent to online social networks. Our results showed that strong correlations are hard to find between these features and churn amongst the entire population of forums. Users are normally unequally established in the community and hence different factors can influence their churn differently. One simple approach is to filter out any users with less than a certain number of contributions during a given window of activity [5]. A more complex approach is to place users in different behaviour categories (e.g. leaders, contributors, followers, grunts) [2] and find the features that correlate best with their churning actions.

The activity and churn windows we used were 13 weeks each, to ensure that sufficient time is given for new activities to emerge before measuring churn. This window size was shown to give best results [5]. However, in low activity forums, such as forum 224, several months could sometimes pass before a new post is contributed and a forum member resurfacing to respond. Therefore, the window size should probably be tied to the level of activity in a forum. Besides the issue of window size, Table 3 shows that higher drops in activity ($\sigma = 0.7$) give more accurate churning predictions (higher F_1 value) to forums with higher activity levels, and vice versa. In 2006, the average number of posts per month for forums 7, 524, 512, and 224 were 14.344, 251, 65, and 9 posts respectively. Hence, a $\sigma = 0.2$ in low activity forums could be much more significant than in very active ones.

Our analysis showed that the centrality of users play a notable role with regard to churning. The feature of Centrality was ranked first in three of our forums, and second in the fourth forum (see Table 4). Figure 3 shows a high churning probability for 'central' users, which could be because these users are more exposed to, and thus aware of, the withdrawal of others from the community. It is also interesting to see that network features (in-degree, out-degree, betweenness, centrality) play a more important role in churning in busy forums (e.g. forum 7) than in others. This could simply be due to having a dense network graph where such features are more likely to thrive. Also, people in large forum communities are more likely to be drawn towards network 'celebrities' and 'leaders', and hence their churn becomes more sensitive to these features than to users in more quite and smaller forum communities. On the other hand, when activity in a forum is low, churn of members can be more damaging to the social network and content flow, and hence it is more likely to push their neighbours to also churn (Figure 4).

6. CONCLUSION

In this paper, we have analysed the relation between user value – characterised as a variety of behavioural and structural user features – and the likelihood that a user will churn, i.e. leave a community. Global analysis identified correlations between lower in-degree and out-degree levels and churning behaviour, while our analysis of sub-communities found differing behavioural patterns unique to certain forums. We also identified network effects, suggesting that the behaviour of an individual's peers effects the probability that the user will remain in the community.

There is a wealth of directions that we identified for future work. We will focus on exploring the reasons for the existence of correlations. This will involve characterisation of the forums analysed here using several statistics, not only but also based on activity levels. In addition to the extended approaches mentioned above, we plan to apply our analysis to several collections of forums with similar characteristics. This will enable us to investigate churn correlations in isolation of any impact caused by, for instance, variant community activity levels. Further, we plan an exhaustive evaluation of network effects. For this, we will analyse the features of a user's neighbourhood and relate it to the individual churn probability.

Acknowledgements

This work was partly funded by the European Union (EU) and Science Foundation Ireland (SFI) in the course of the projects ROBUST (EU grant no. 257859), WeGov (EU grant no. 248512), the CLIQUE Strategic Research Cluster (SFI grant no. 08/SRC/I1407), and LION-2 (SFI grant no. SFI/08/CE/I1380).

7. REFERENCES

[1] J.-H. Ahn, S.-P. Han, and Y.-S. Lee. Customer churn analysis: Churn determinants and mediation effects of partial defection in the korean mobile telecommunications service industry. *Telecommunications Policy*, 30(10-11):552 – 568, 2006.

[2] J. Chan, E. M. Daly, and C. Hayes. Decomposing discussion forums and boards using user roles. In *ICWSM'10*, pages 215–218, 2010.

[3] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi. Social ties and their relevance to churn in mobile telecom networks. In *EDBT '08*, pages 668–677, 2008.

[4] P. Dwyer. Measuring the value of electronic word of mouth and its impact in consumer communities. *Journal of Interactive Marketing*, 21(2):63–79, 2007.

[5] M. Karnstedt, T. Hennessy, J. Chan, P. Basuchowdhuri, C. Hayes, and T. Strufe. Churn in social networks. In B. Furht, editor, *Handbook of Social Network Technologies and Applications*, pages 185–220. Springer Verlag, 2010.

[6] J. Kawale, A. Pal, and J. Srivastava. Churn Prediction in MMORPGs: A Social Influence Based Approach. In *CSE '09*, pages 423–428, 2009.

[7] P. Kollock. *The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace*. Routledge, London, 1999.

[8] M. Mozer, R. Wolniewicz, D. Grimes, E. Johnson, and H. Kaushansky. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, 11(3):690–696, May 2000.

[9] J. J. Phillips. Human capital measurement: A challenge for the clo. Technical report, Clo Media, 2003.

[10] C. Shirky. Broadcast institutions, community values, 2004. `http://www.shirky.com/writings/broadcast_and_community.html`.