

Churn in Social Networks: A Discussion Boards Case Study

Marcel Karnstedt, Tara Hennessy, Jeffrey Chan, Conor Hayes
Digital Enterprise Research Institute (DERI), NUI Galway, Ireland
Email: *firstname.lastname@deri.org*

Abstract—Churn has been identified as an important issue in a wide range of industries. In social networks, churn represents a significant risk for the health and functioning of communities. However, the importance and actual meaning of churn in social networks is almost unexplored. This work provides a general view on these issues and discusses aspects that are especially relevant to discussion boards. We provide a broad literature review on “traditional” churn analysis and prediction and highlight the specialities of churn in social networks. We further present an empirical analysis of a churn definition particularly appropriate for discussion boards and propose future research directions for predicting churn in social networks, focusing on the importance of social roles, influence and influence diffusion.

I. INTRODUCTION

In the past, churn has been identified as an issue across most industry sectors. In its most general sense it refers to the rate of loss of customers from a company’s customer base. There is a simple reason for the attention churn attracts: churning customers mean a loss of revenue. Emerging from business spaces like telecommunications (telcom) and broadcast providers, where churn is a major issue, it is also regarded as a crucial problem in many other businesses, such as online games creators, but also online social networks and discussion sites. Companies aim at identifying the risk of churn in its early stages, as it is usually much cheaper to retain a customer than to try to win him or her back. If this risk can be accurately predicted, marketing departments can target customers efficiently with tailored incentives to prevent them from leaving.

While churn in the telcom sector has been studied extensively, no research has been carried out on the general meaning and consequence of churn in online social networks and communities. Churn is an important factor for social network providers, as it leads to a loss of social capital and ultimately effects service sustainability. Churn undermines the health and ability of communities to self-govern and self-maintain. The risks to digital social networks arise not only through members stopping their activity, but also through significantly reducing activity. This becomes particularly evident where network services rely upon member activity, such as in discussion boards. Often, the combination of increasing and decreasing activity (of different customers) is an important indicator for risk. As an example, consider a forum where the popular topic initiators are gradually replaced by spammers. Such a change in the population composition indicates the risk of “losing” the forum, its associated services and revenue

model. Therefore, churn analysis should be understood in close relation to analysing activity.

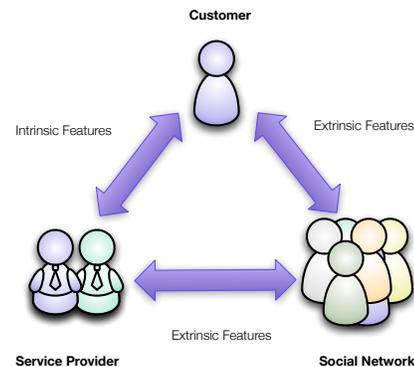


Fig. 1. Intrinsic and extrinsic features of churn

Factors contributing to churn have intrinsic and extrinsic features, see Figure 1. Intrinsic features relate to inherent qualities of the service/product and/or service provider. While much research has been carried out on how intrinsic features can be used to predict potential churners, there is comparatively little research on how extrinsic, social features contribute to churn. In this work, we explore this idea in greater detail.

In this work, we focus on the differences between churn in traditional areas and digital social networks, specifically discussion boards. We first present a broad overview of related work in Section II. Then, we discuss the special aspects of churn in digital social networks and present the reasons and motivations underlying user activity in Section III. We propose a definition of churn for discussion boards in Section IV and show its applicability on the basis of an empirical study in Section V. Finally, in Section VI, we briefly discuss the importance of social roles, influence and diffusion models for churn prediction. We provide implications for further research in order to leverage novel diffusion-based methods for churn prediction by using social roles, and discuss how to combine these methods with traditional feature-based approaches. The work is concluded in Section VII.

II. CHURN IN THE LITERATURE

Churn has been analysed in a wide range of industries: most widely in the telcom sector [1], [2], [3], but also in the fields of banking [4], Internet service providers [5], P2P networks [6], online games [7], and much more. In its most general sense churn refers to customer loss. In the telcom

industry, a subscriber is said to have churned when he leaves one carrier to move to another [3]. Churn rate is defined as the total gross number of subscribers who leave the service in the period divided by the average total customers in the period. The churn rate of a telcom company is a key measure of risk and uncertainty in the marketplace and will be quoted in the company annual report.

Keaveney gives several factors influencing churn in the service industry such as pricing, inconvenience, core service failures, customer service failures and dissatisfaction with provider ethics [8]. In subsequent studies, factors such as these have been decomposed into feature profiles of the customer and customer experience in order to be able to predict churn [3]. Consequently, churn prediction is usually based on pure *feature-based* approaches [9]. However, this neglects the importance of social influence between individuals, which can be analysed by examining their social networks. We believe that churn is not only a phenomenon due to individual decisions and profiles, rather it is influenced by external events (e.g., elections or poor reviews for a company) and, more importantly, by community effects. Extrinsic features describe the service in terms of the value it accrues through its social role. This intuition is captured by Metcalfe's "law", which states that the value of a communications network is a function of the number of connected users of the system [10]. Telcom providers that offer discounted calls to the subscribers' family and friends make use of this effect to tie the service to the stability and longevity of a customer's real social network. A related extrinsic function is where the value of a network service is a function of the *opinion* of the customer's social call network. Although there is not a "law" to describe this function, *word-of-mouth* algorithms that combine neighbour preferences to predict the preference of a target user are well established in the domain of online recommender systems [11]. Such algorithms explicitly recognise that extrinsic, social features are as important as intrinsic features in determining the perceived value of an object or service.

To analyse and understand community effects, a crucial aspect is to understand social roles of single individuals and their influence on the community. Some recent work has identified this issue and argues to replace or extend the feature-based approaches by *influence-based* techniques, mostly by the use of information diffusion models [2], [7]. The idea underlying this approach is that a subscriber is more likely to churn if he is connected to other subscribers that have churned. Thus churn diffuses from subscriber to subscriber, where the degree of influence of the previous churners depends on their importance and social weight. We believe that this is the right approach, but that the applied *spreading activation* (SPA) model [12] bears several problems and that diffusion-based approaches should be combined with feature-based approaches. We discuss this further in Section VI.

In P2P networks, churn is defined as the collective effect of the independent join-participate-leave cycles of all peers observed over a given period [6]. Churn in P2P networks

affects key design parameters and the resiliency and structure of the overlay network. While this introduces the notion of cyclic join-participate-leave behaviour in low entry cost dynamic systems, P2P churn differs from telcom and social network churn because there are no social or contractual ties affecting participation, i.e., peer behaviour is independent.

Several studies suggest that retaining existing customers is considerably less expensive than winning new customers [1], and that new customers tend to be less profitable than existing customers [13]. Lifetime value (LTV) analysis is often used to predict the future profitability of a potential churning so that only the most valuable subscribers are targeted for retention [3]. In P2P simulation analysis, Herrera and Znati [6] introduced the concept of different classes of peers, some of whose behaviours play a crucial role in the applications that the network can support. The idea of segmenting contributors in social networks into different value categories based on their predicted value to the community has only begun to be explored [14]. The authors propose that the presence of roles such as 'popular initiator' or 'supporter' are important for the health and sustainability of the network [14]. In Section VI, we begin to develop this idea further.

III. UNDERSTANDING CHURN IN SOCIAL NETWORKS

The value that subscribers derive from activity in online digital networks is tied directly to the social capital of the network. The loss of capital through subscriber loss or significant drop in activity may precipitate spiralling cycles of churn. 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 [15]. In sociological discourse, this type of activity is described in terms of the 'gift economy' [16]. In contrast to the commodity or service economy, which is driven by the exchange of good/services for money, economic exchange in the gift economy is defined in terms of an implicit social contract. In a gift transaction, there is an unstated expectation that the benefits of a gift will be reciprocated by the recipient at some reasonable time in the future. A more risky transaction involves 'generalised exchange', whereby the giver's generosity is reciprocated, not by the recipient, but by someone else in the group.

In social networks, this exchange mechanism applies to those contributors who give of their time and expertise but do not appear to receive immediate benefits. However, there is a risk that the group will not assume responsibility for the debt and the contributor will never be reimbursed in kind. In the worst case, if all members of the group never contribute (free-load), no one benefits and the exchange system breaks down. Despite this risk, several motivations for contribution to digital social networks have been proposed, e.g., in [15]. One such is the contributor's expectation of reciprocity under the generalised exchange system. Contributors expect that there will be sufficient payback in terms of information and agreeable social relations from other contributors over time. Another motivation is the contributor's desire to enhance his

reputation and prestige by being recognised as a source of valuable information or help. The contributor may also be motivated by trying to have his/her ideas adopted by others in the group, or by the value he/she places in the attachment to the group or the values represented by the group. Thus, contribution is partly made to sustain the group and the contributor’s continuing attachment. For each motivation, the contributor derives value in part from his contribution in a social context. We can hypothesise that churn is an outcome that tends to occur when the expected return drops below a certain threshold - perhaps the cost of contribution. In short, we hypothesise that the contributor’s utility function is dependent on the social characteristics of the network.

The subscriber to an online community is weakly-tied by a non-binding social contract [15]. The telcom subscriber is generally bound by a service contract or by the purchase of advance credit. In general, subscribers have a contract with one telcom provider and tend to break that contract in response to industry-understood ‘triggering events’, such as the expiry of the service contract or a poor customer service experience [17]. The main prohibition to customer churn is the switching cost involved in moving to another service provider, such as the loss of loyalty points [18]. Online communities do not have a switching costs. There is a low-entry barrier to join and the contributor may leave without incurring an explicit penalty. More importantly, the contributor may return at any time, and may have simultaneous involvement in several communities. As such, churn is defined by telcom providers in terms of complete loss of custom, while in online communities the notion of churn is less easily defined.

In fact, the ease with which contributors can alter their behaviour online suggests that, in contrast to the telcom view, churn is a normative behaviour in social networks and that there will always be an underlying turnover in contributor activity. As such, churn in social networks refers to turnover where there is a net loss of the types of contributors that are needed to maintain the service level of the network. Therefore, churn can be analysed on different levels: per user, per thread, per forum, per community, etc. Unlike in telcom networks where churn is defined in terms of a customer’s complete defection, a contributor in a social network may be deemed to be a ‘partial defective’ if his activity drops significantly below previously established levels of engagement. The overall effect of churn on a social network may be a function of the number of partial defectives who are also high value contributors. By moving from a binary decision (activity stopped or not) to one based on a significant drop in activity, the question of an appropriate churn definition arises. We discuss this in the next section. Moreover, as we show as well, this results in the existence of different types of churn, depending on the significance and duration of drop in activity – and the actual risks and defectives that shall be identified.

IV. DEFINING CHURN FOR DISCUSSION BOARDS

As discussed in Section III, the low cost in contributing and in exiting means that online contributors demonstrate

behaviour that cannot easily be categorised in terms of binary churn. In this section, we propose an intuitive definition for churn and several variants, while indicating the implications and suitability of different alternatives for several aims and domains. In its most general sense churn is perceived as a significant and sustainable change in the activity of single individuals and/or communities. We argue that no general definition exists that would be the best for each case. Rather, the chosen definition should depend on the application domain and the risks that churn analysis and prediction should indicate. This section is not meant to provide an exhaustive discussion of possible definitions. Rather, its main purpose is to highlight the crucial points for choosing such a definition.

Definition 1: Individual Churn: The *previous activity (PA)* window is a time window consisting of time steps t_1 to $t_1 + n - 1$ inclusive, $n \in \mathbb{N}, n \geq 0$. Let $\mu_{PA}(v_i)$ denote the average activity of a user v_i over the previous activity window. The *churn (C)* window is a time window $t_2 = t_1 + n$ to $t_2 + m - 1$ inclusive, $m \in \mathbb{N}, m \geq 0$. Let $\mu_C(v_i)$ denote the average activity of a user v_i over the churn window. A user v_i is considered to have churned during the churn window if:

$$\mu_C(v_i) \leq T(\mathcal{S}) \cdot \mu_{PA}(v_i)$$

$0 \leq T(\mathcal{S}) < 1$ is a threshold factor dependent on the relevant system parameters \mathcal{S} .

The definition is more simply stated as:

A user has churned if his or her average activity over a window of m time steps has dropped to less than a fraction T of their average activity in the previous n time steps, where m and n are positive integers and $T < 1$ is a suitable system-dependent threshold factor.

In Section V, we investigate the sensitivity of the above definition to varying parameters n, m and T . Note that it is also possible to define the threshold factor dependent on the user him- or herself, i.e., use v_i as an additional parameter for T . The identification of a suitable threshold factor, the system parameters it depends on and the appropriate past activity present particularly interesting problems. Several variants of Definition 1 can be differentiated based on these aspects. One alternative is to make churn dependent on whatever constitutes typical activity of the average user in the entire system.

Definition 2: Mean activity: Let $a(v_i, t)$ denote the activity of a user v_i at time t , N the total number of time steps in the observations and V denote the set of all users of the network. The mean activity μ_u across all users is defined as:

$$\mu_u = \frac{1}{N \cdot |V|} \sum_{v_i \in V} \sum_{t=1}^N a(v_i, t)$$

Based on this mean activity, the following variants of Definition 1 are intuitive:

- **Global Churn:** Replace $\mu_{PA}(v_i)$ in Definition 1 by μ_u , use a global threshold factor defined as some function $T(\mu_u, \mathcal{S})$, or both.
- **Role Churn:** Define separate $\mu_R(r)$ for each existing role r . The definition of $\mu_R(r)$ follows straightforwardly from

Definition 2 by restricting to only users with role r . Then, replace $\mu_{PA}(v_i)$ in Definition 1 by $\mu_R(r)$, use a role threshold factor defined as some function $T(\mu_R(r), \mathcal{S})$, or both.

Roles can be found in a wide range of social networks. For instance, users of forum sites can be classified by the common behaviour roles they play [14]. Some users could play the *popular initiator* role, i.e., tend to initiate many threads and get many replies to their posts. In contrast, other users can play the role of *taciturn*, i.e., have low posting and replying behaviour. The normal and abnormal churning behaviours of the two sets of users are going to be different. Hence, it makes sense to define churn on the basis of the importance of a user, i.e., on the basis of the role they play. Similarly, this can be extended to other levels, dependent on the domain of the system. For instance, for discussion boards, it might be interesting to inspect the typical activity in each forum:

- **Forum Churn:** Define separate $\mu_F(f)$ for each existing forum f . The definition of $\mu_F(f)$ follows straightforwardly from Definition 2 by restricting to only users active in forum f . Then, replace $\mu_{PA}(v_i)$ in Definition 1 by $\mu_F(f)$, use a forum threshold factor defined as some function $T(\mu_F(f), \mathcal{S})$, or both.

All the above variants of Definition 1 are only examples of how to define churn. Some obvious possibilities to further adapt them, without changing their general meaning, are to replace the mean by median, to choose other relative thresholds or to choose absolute thresholds. Alternatively we could require the activity levels to be below the threshold for a continuous span of time, rather than on average.

For certain reasons it might be interesting to identify a user as churning as soon as his activity drops below the threshold for the first time. This can be achieved with Definition 1 by setting $m = 1$. As introduced in Section III, in social networks different types of churn can occur on different levels. The definition is general and flexible enough to allow several different types of churn and related phenomena to be identified and analysed. Examples of such types for discussion boards are:

- Detect the emergence of a new forum population: Forum Churn for existing users and increasing activity of other users in the same forum.
- Detect movement of activity from one forum to another: Forum Churn in one forum and increasing activity of the same users in other forums.
- Detect the change of the role of an individual user: Role churn of the user while staying above the threshold for another role.
- Churn on different levels of an application can always be mapped to the churn of single individuals: a “dying” discussion thread can be identified if most of the posting users are classified as churners, a “dying” forum is composed of dying threads, etc.

We further investigate the sensitivity of the definition with respect to the included parameters and resulting types of

identified churn in Section V. Other definitions for churn may be based on the degree of change in activity, i.e., a point in time when the magnitude of the decreasing rate of change of activity of a user is above some absolute threshold, the ratio between variance and mean of activity, change point detection, etc. However, the discussions presented here highlight the crucial parts in the definition of churn: the applied threshold (factor) T , the window sizes for past activity and churn window as well as the type of past activity used. The right choice is dependent on the domain and the risks that need to be identified. As an example, consider the telcom and online game sectors. To identify likely churners, it is sufficient to identify customers that will drop to an absolute threshold of $T = 0$. The earlier we require to have potential churners identified, the higher this absolute threshold should be. This is based on the observation that usually activity slightly decreases in the months before the decision to churn [7]. The gradient churn definition may be useful to detect churn in any state of the customer lifecycle, rather than only shortly before the churn event. On the other hand, as resources for customer care are usually restricted, providers might want to focus on high-value customers. In this case, applying Role Churn might be the right choice.

A different picture is drawn in social networks that do not imply churning costs for customers, such as discussion boards. Here, risks related to churn materialise in different states and should usually be classified according to the characteristics of the single user, such as his role in a forum. Obviously, relative thresholds for Individual or Role Churn are an intuitive choice in this case. Even if the activity slightly increases after a significant drop, but stays below the critical threshold, this should be understood as an alarming signal in such domains. Whereas in the telcom space it implies that a customer is still active in the network of the operator. In the following section, we present a brief empirical analysis to highlight the effects of different parameters for Definition 1 in such social networks.

V. EMPIRICAL ANALYSIS

Identification of churn in a network will critically depend on whatever constitutes typical behaviour for a user in that network. In this section, we present a basic empirical analysis using a data set gained from a discussion board. We indicate the sensitivity of Definition 1 from Section IV with respect to the choice of the window sizes m and n , and the threshold factor T . We further show that with different parameter combinations it is possible to isolate significantly different types of churn. First, we briefly introduce the data set used.

A. The Used Data

The data set consists of the number of posts made per week by individual users in individual forums of the popular Irish forum site <http://boards.ie> over the course of the year 2006. In the following, we refer to the number of posts vs. time as *activity profile*. On initial inspection of the data the activity profiles for single users appear very discontinuous and “jagged”. The data set, in its entirety, consists of almost

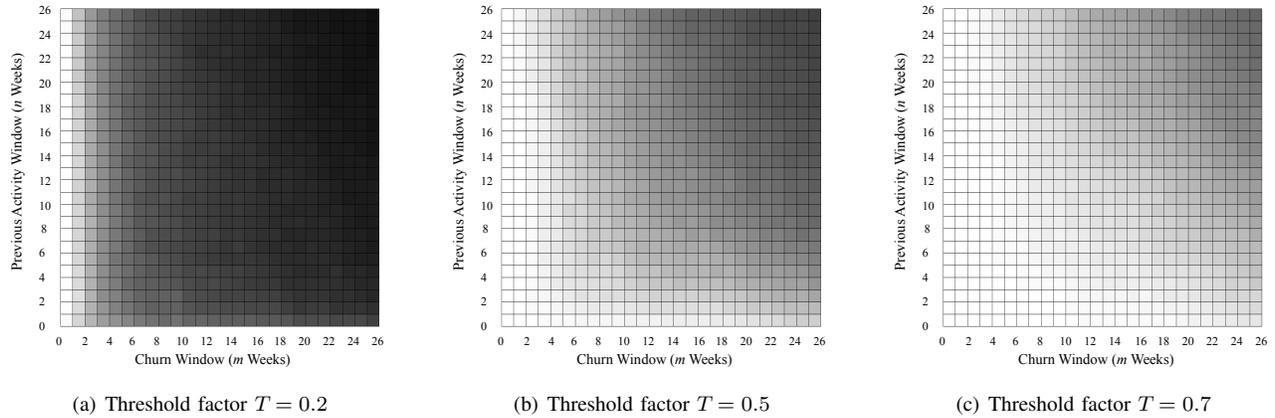


Fig. 2. The number of users identified as churners (from black=0% to white=100%) is very sensitive to the chosen parameters: previous activity window size n , churn window size m and threshold factor T .

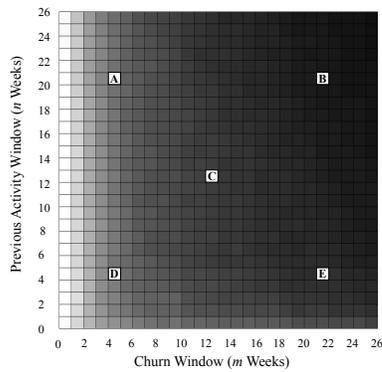


Fig. 3. The five parameter sets chosen for further investigation.

150,000 activity profiles. However, a large percentage of the users make 1-2 posts and never return. These users are of little interest for churn analysis. The loss of regular users is of much greater concern to a service provider. Thus, the data was filtered to isolate users who made an average of at least 10 posts per week. The remaining sample had a size of 300 users. The number of users identified as churners in any window over the course of 2006 was recorded for a range of window sizes and threshold factors. A maximal churn window size m of 26 weeks is a natural choice, as it is half of the total time spanned by the full data set investigated. We applied Definition 1 to each activity profile. Selected results of this analysis are shown in Figure 2. The individual figures correspond to different threshold factors. The colour gradient ranges from black, when nobody churns, to white, when all of the sample of 300 are identified as churners.

In Figure 2(a) it can be seen that when the threshold factor is very low (20%), the condition for churn is quite strong. Thus, for large window sizes m and n very few users are identified as churners. For smaller window sizes more churners are identified. This is due to the nature of the data, as almost all users have a short drop of activity at some point in time. This explains the white column at the far left hand side and shows that such a small window size is not appropriate to identify “real” churn. Along this line the churn window is only one week long, which means that for churn to be identified the

user has to be away from the site or have uncharacteristically low activity for only one week. In Figures 2(b) and 2(c) it can be seen that as the threshold is raised the conditions for churn become less strict. The number of users identified as churners raises accordingly.

Parameter Set	Previous Activity Window size n	Churn Criteria Window size m	No. of churners (out of 300)
<i>A</i>	21	5	130
<i>B</i>	21	21	32
<i>C</i>	13	13	69
<i>D</i>	5	5	137
<i>E</i>	5	21	53

TABLE I
NUMBER OF CHURNERS IDENTIFIED BY DIFFERENT WINDOW SIZES.

In Section III, we mentioned the importance of different types of churn that can occur in social networks. In the following, we present a more detailed analysis using a threshold factor of 20%. Regarding the number of identified churners for different window sizes, it is crucial to understand the relations between parameter values and the according types of identified churn. The threshold factor of 20% represents a rather strict condition and is in-line with an intuitive understanding of churn.



Fig. 4. The approximate hierarchy.

1) *An Approximate Hierarchy*: To analyse the effects of parameter choices in more detail, the parameter combinations at *A*, *B*, *C*, *D* and *E* in Figure 3 were chosen. They result in significantly contrasting variants of the definition and thus provide insights into the different types of churn that are identified.

As the parameter choices vary from the bottom left of Figure 3 in any positive direction (i.e., as either of the window sizes increases), the criteria for identifying churn becomes stricter. This holds not only along the diagonal, but also,

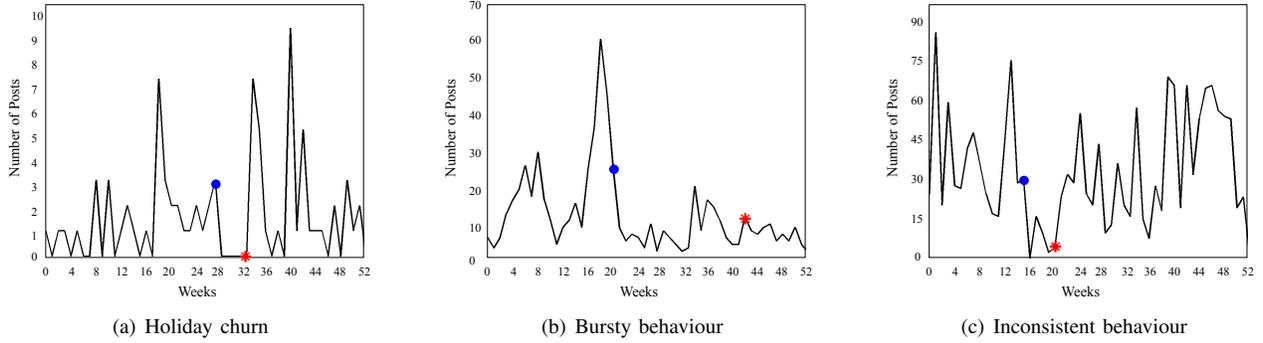


Fig. 5. Examples for different types of churn from the approximate hierarchy.

for instance, as we move from parameter combination D to combination A or E . Table I supports this observation by showing the numbers of users identified as churners for the different parameter combinations. The different combinations seem to form a hierarchy, meaning that, for example, all churners identified in A are also identified in D . In fact, this is only an approximate hierarchy, illustrated in Figure 4.

Now, the question is what different types of churn are identified by the different parameter combinations. Some interesting examples of churn are found through identifying exceptions to the hierarchy. We show examples for these exceptions in Figure 5. In each of the figures, the circles represent the beginning of a churn window in which churn is identified and the stars represent the end of the window. Due to the definition, there may be multiple consecutive churn windows for which churn is detected. The circle and star corresponding to the first churn window are shown in bold.

Typical churn: churn identified by parameter combination B: Criteria B can be seen as the strictest criteria of all five. It requires that the average activity of the user drops significantly for the longest time, while having been high for an equally long time. The users identified by B are those that we see as *typical churners*. An example is shown in Figure 6. Out of the 32 churners identified by criteria B , all but 2 are also identified by D and E , all but one are also identified by C and all 32 are identified by A .

Holiday churn: churn identified by parameter combination A, but not by B or D: We found examples of churn that are identified by parameter combination A , but not by combinations B and D . Such examples might be called *holiday churn*, where a pattern of consistently high activity is interrupted for a short period of 2-4 weeks. Such short drops in activity have to be pronounced enough to significantly affect the average activity when they fall into a short churn window. But, they do not significantly affect the average when they represent only a small part of a much longer churn window. We show an according example activity profile in Figure 5(a).

Bursty behaviour: churn identified by parameter combination E only: E is the parameter combination that refers least to an intuitive understanding of churn. It tends to identify short, uncharacteristically high periods of activity that are preceded by long periods of low or even just average behaviour. This is due to the short previous activity window and the long churn

window. We refer to this as *bursty behaviour*. An example, which is identified by E but not by any other combination, is shown in figure Figure 5(b).

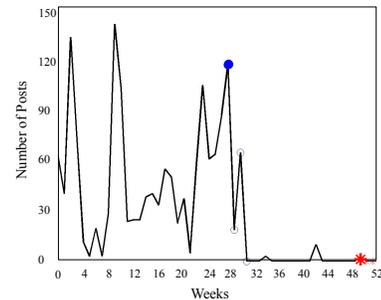


Fig. 6. Typical churn behaviour.

Inconsistent behaviour: churn identified by parameter combination D only: Combination D imposes the least strict criteria for churn. Most users have some uncharacteristically low period of activity for some weeks in the year. Since criteria D has two short windows, it flags a kind of *inconsistent behaviour* as churn, see Figure 5(c) for an example. By inspecting the amount of churn identified by combination D but not by any other combination, one could try to measure the consistency of users.

Combination C: Combination C , right in the centre of Figure 3, represents a balance between all other combinations. It works well for identifying typical churn, as well as detecting significant periods of churn before recovery. Due to this generality, it is the combination that we use in Section VI to analyse network effects for churn.

B. Summary

By applying Definition 1 to the data, we have shown that the amount and variety of churn identified is highly sensitive to parameter choices. Decreasing the size n of the past activity window increases the sensitivity of the churn detection, i.e., low activity for shorter periods is more often understood as churn. Consequently, increasing n results in detecting more significant churn in an intuitive sense. Increasing the size m of the churn window obviously increases the time until churn is detected as well as the duration of periods of low activity required for indication. Further, we have shown that some

interesting types of churn can be isolated by using various combinations of parameters.

The brief empirical analysis presented here shows that the proposed definition for churn is suited for a wide range of different social networks. Depending on the actual motivation and needs of the analyser. The risks and phenomena that can be identified, from typical churn, over bursty and (in-)consistent behaviour, to sensitively recognised fluctuations in activity. However, other definitions for churn, as well as data from different kinds of social networks, should be analysed and compared in future work. The presented discussion highlights the importance of the correct understanding and modelling of churn, and the significance of carefully investigating this issue before any meaningful prediction or prevention mechanisms can be developed.

VI. NETWORK EFFECTS AND PROPAGATION OF CHURN

Discussion boards like boards.ie do not always define an explicit network structure. Different influence network views can be built based on connections between users that refer to replies, activity in the same forum or thread, etc. One useful network view is the user-to-user interaction network, for instance, based on the reply-to structure (nodes represent users, an edge represents one user replying to another in a forum thread). In this section, we show that there are network effects for churning based on this interaction network. The idea is to infer the influence of one user on another by the amount of communication between them.

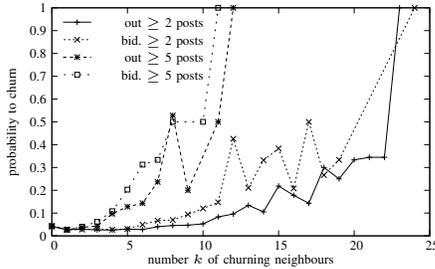


Fig. 7. Probability to churn

Similar to preliminary experiments in [2], [7], Figure 7 shows an individual's probability to churn in relation to the number k of the neighbours that already churned. This probability is determined by dividing the number of all current churners that have k churned neighbours by the total number of users that have k churned neighbours. In other words, it is the percentage of all people with k churning neighbours who decided to churn as well. The figure shows this relation for four different reply graphs. The number of replies constitutes the weight of an edge. We averaged the values over all forums and over all analysed time slots. The labels with `out` refer to the graph based on only outgoing edges, `bid.` refers to the graph containing only bidirectional edges (thus, it is sparser). To highlight the effects of different influence, we filtered edges by their weight: ≥ 2 contains only edges with a weight larger than 1, ≥ 5 contains only edges with weight larger than 4.

As expected, the churn probability increases with k , which illustrates the existence of network effects (i.e., the more

churning neighbours there are, the more likely a user is to churn also). The fact that this probability is 1.0 for some k is because only a few users have that many churning neighbours, and all these users churned as well. To illustrate this, we show the total number of churners with k churned neighbours in Figure 8. Note that the y axis is shown in log scale. Comparing the outgoing edge results for ≥ 2 and ≥ 5 posts in Figure 7 shows that the amount of influence is apparently increasing with the number of replies between users.

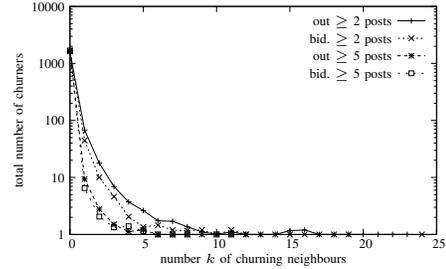


Fig. 8. Total number of churners

In previous work [14], the authors grouped the common features of users in boards.ie and proposed 8 different social roles played by users in boards.ie. These were: joining conversationalists, popular initiators, taciturns, supporters, elitists, popular participates, grunts, and ignored. The roles were determined on the basis of conventional and structural features, based on the reply-to interaction network. Thus, the network effects presented above relate to the roles determined in [14]. We therefore posit that different roles have different amount of influence on other roles. For example, users playing the elitist role will have strong influence on other users playing the same role, but users in all other roles will have little or no influence on the users of the elitist role. One value or measure of influence is hence not appropriate. We propose that we need to examine conditional influence - conditional on the users and possibly features. There has been some recent work in this direction for determining different influence measures for different topics [19] and movies [20], which we plan to explore for determining relative influence of roles.

Only recently, some first works [2], [7] consider social network analysis as an alternative or extension to customer churn prediction models solely based on features. In both works, the application of the SPA model helped to increase the precision of prediction accuracy significantly. SPA can be applied easily to many applications. However, it is a rather simple model for information diffusion. Tuning global parameters poses a problem, as the model does not differentiate between different roles and influence of individuals. We believe that some of the well-known diffusion models are better suited to predict churn based on the influence of individuals than the introduced SPA model. [21] provides a broad survey and further pointers to more detailed surveys for most of the traditional models. For an improved churn prediction, these models should be combined with positive and negative influences, e.g., by combining linear thresholds with a sentiment measure. Furthermore, similarity in an user's

and neighbour's features (homophily) and its effect on churn diffusion should be investigated. The crucial point for an appropriate application of diffusion models for churn prediction will be the inclusion of social roles, popularity and the according individual (conditional) influence of users into the used diffusion model. In addition, diffusion-based models for predicting churn should take external circumstances into account. This should include external events (e.g., a new provider enters the arena with special offers), the importance of providers' pricing politics and reputation to geographical aspects. A last aspect that up to now did not receive enough attention is to analyse the "direction" of influence, i.e., is the influence implicit or is it explicitly initiated by the churning.

Intuitively, one should not have to decide between either a feature-based or a diffusion-based approach. Rather, both ideas should be combined accordingly to achieve prediction accuracy and efficiency that none of the two approaches could achieve on their own. To the best of our knowledge, up to now there is no concrete suggestion or proposal of how to exactly combine both worlds. Further research has to analyse the possibilities of using features to adapt the influence of individuals effective in the diffusion process. We propose to focus further research on an enrichment of the *conventional* features by *structural* features (connectivity, degree, triangles, path lengths, etc.). An alternative direction worth investigation is the use of an inverse approach, i.e., use information diffusion to adapt (as input for) the feature-based approach. However, this direction seems to be not as obvious.

VII. CONCLUSION

Research on churn currently enjoys great popularity due to the loss of revenue to service providers in which it results. In the literature to date, the focus usually lies on contractual services like in the telcom sector. However, the notion of churn and the factors driving it, such as the social costs compared to purely monetary costs, and the risks in social networks have not been widely investigated. This paper aims to fill this gap. In this context, we defined churn as an act of significantly decreasing activity. We further highlighted the importance of being able to predict churn in social networks, as it relates to the health and functioning of the underlying communities.

Firstly, we have discussed different notions, reasons and facets of churn in social networks compared to other domains. We have further given an overview of current research on churn detection and interpretation. After providing our own suggestion for defining churn in discussion boards, we demonstrated its effectiveness and crucial parameters on the basis of a large dataset from boards.ie.

The results presented here show that the proposed churn definition is appropriate for identifying a wide range of different churn types that are present in social networks. This enables the detection and prediction of churn in social networks for a wide range of applications and potential risks. Further, it highlights the crucial factors in understanding and handling churn. We showed that there is a relationship between churn, network effects, social roles and influence. We proposed the

combination of traditional feature-based churn prediction with diffusion models and highlighted possible directions and open issues for this novel field of research in social networks.

ACKNOWLEDGMENT

This work was carried out in part in the CLIQUE Strategic Research Cluster, which is funded by Science Foundation Ireland (SFI) under grant number 08/SRC/I1407.

REFERENCES

- [1] M. Mozer, R. H. Wolniewicz, D. B. Grimes, E. Johnson, and H. Kaushansky, "Churn reduction in the wireless industry," in *NIPS*, 1999, pp. 935–941.
- [2] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjee, A. A. Nanavati, and A. Joshi, "Social ties and their relevance to churn in mobile telecom networks," in *EDBT '08*, 2008, pp. 668–677.
- [3] H. Hwang, T. Jung, and E. Suh, "An ltv model and customer segmentation based on customer value: a case study on the wireless telecommunication industry," *Expert Syst. Appl.*, vol. 26, no. 2, pp. 181–188, 2004.
- [4] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5445–5449, 2009.
- [5] B. Q. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction for broadband internet services," in *DaWaK*, ser. Lecture Notes in Computer Science, vol. 5691, 2009, pp. 229–243.
- [6] O. Herrera and T. Znati, "Modeling churn in P2P networks," in *Annual Simulation Symposium*. IEEE Computer Society, 2007, pp. 33–40.
- [7] J. Kawale, A. Pal, and J. Srivastava, "Churn Prediction in MMORPGs: A Social Influence Based Approach," in *CSE '09*, 2009, pp. 423–428.
- [8] S. M. Keaveney, "Customer switching behavior in service industries: An exploratory study," *The Journal of Marketing*, vol. 59, no. 2, pp. 71–82, 1995. [Online]. Available: <http://www.jstor.org/stable/1252074>
- [9] J. Burez and D. V. den Poel, "Handling class imbalance in customer churn prediction," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4626–4636, 2009.
- [10] B. Metcalfe, "Metcalfe's law: A network becomes more valuable as it reaches more users," 1995.
- [11] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *J. Mach. Learn. Res.*, vol. 4, pp. 933–969, 2003.
- [12] G. Salton and C. Buckley, "On the use of spreading activation methods in automatic information," in *SIGIR '88*, 1988, pp. 147–160.
- [13] W. Buckinx and D. V. den Poel, "Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," *European Journal of Operational Research*, vol. 164, no. 1, pp. 252–268, 2005.
- [14] J. Chan, E. M. Daly, and C. Hayes, "Decomposing discussion forums and boards using user roles," in *AAAI Conference on Weblogs and Social Media*, 2010, to appear.
- [15] D. Constant, L. Sproull, and S. Kiesler, "The kindness of strangers: The usefulness of electronic weak ties for technical advice," *Organization Science*, vol. 7, no. 2, pp. 119–135, 1996.
- [16] H. Rheingold, *The Virtual Community: Homesteading on the Electronic Frontier*. Cambridge, MA, USA: MIT Press, 2000.
- [17] A. Gustafsson, M. Johnson, and I. Roos, "The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention," *Journal of Marketing*, vol. 69, no. 4, pp. 210–218, 2005.
- [18] J. P. Maicas, Y. Polo, and F. J. Sese, "Reducing the level of switching costs in mobile communications: The case of mobile number portability," *Telecommunications Policy*, vol. 33, no. 9, pp. 544 – 554, 2009. [Online]. Available: <http://www.sciencedirect.com/science/article/B6VCC-4WCT02M-1/2/737ee54afa660372dbd56e74d6fec8c4>
- [19] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *SIGKDD '09*, 2009, pp. 807–816.
- [20] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Discovering leaders from community actions," in *CIKM '08*. New York, NY, USA: ACM, 2008, pp. 499–508.
- [21] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *WWW '04*, 2004, pp. 491–501.