Continued Participation Prediction in Online Social Networks: Challenges and Opportunities

Farig Sadeque
School of Information
University of Arizona
1103 E 2nd St, Tucson, AZ 85721
farig@email.arizona.edu

Abstract

Churn or continued participation has been one of the most researched topics in a wide range of industries. In online social networking paradigms, this phenomenon has largely been unexplored, although having a profound influence on the social capital and sustainability of the networks. This study provides a detailed literature review on churn or continued participation prediction in online social networks along with traditional churn techniques in other industries like telecommunications. The study presents a short history of the transition of traditional churn techniques into the modern participation prediction techniques better suited for social networks, along with the challenges and opportunities in this field of research. It also provides detailed insights into the major approaches that have been implemented for continued participation prediction using user behavior, language, demographics and interpersonal relationships- unraveling their successes and limitations. Special aspects of the social networks determines the approach that is the best fit for its continued participation analysis, and there is a little research dedicating to the adaptability and scalability of these approaches over a wide range of services. This study also proposes future research directions based on the empirically observed limitations of the current approaches, mostly focusing on temporal analysis of user activities.

Introduction

One of the most prevalent use of the Internet nowa-days is the online social networks as they have established themselves as an integral part of human interaction and communication over the last decade. The term "Online Social Network" covers a wide range of services that provide online interaction- from micro-blogging sites such as Twitter to support group based health forums such as DailyStrength. As these networks grew, a lot of research works has been done on them over the years. Yet, one of the most intriguing aspects is mostly overlooked in these research works- continued participation or engagement. Despite the fact that continued participation is one of the most significant contributors of social capital in online forums and social networks, there is little research on computational assessments of the level of engagement in popular social networking sites. Though engagement in large services like Twitter or Usenet newsgroups has been explored over the years, smaller networks, specifically support group or community based forums have received little to no attention.

Continued participation is a frequently researched topic across many industry sectors. The discontinuation in participation in an industry is referred as *Churn*- a shortened version of the combination of two words: *Change* and *turn*. In general sense this word means the rate of loss of customers from a company's customer base to another company. Research on churn has a simple motivation- loss of customer is loss of revenue, and retaining a customer is much cheaper than winning a new one (Hadden et al., 2007). Generally a company tries to identify a churning customer early in their lifecycle so that customer management departments can efficiently target these customers and provide incentives to prevent them from leaving the company. Among these industries, telecommunication sectors have contributed extensively in the research of churn among their customers.

Consequently churn is an important factor to the social network services since they follow the same business model as the service providers in telecommunication sectors- you lose revenue when a customer leaves the network. However, in social networks, the threat is much more than monetary. As all of the social networks thrive on the interactions among users, loss of users means loss of social capital within the service- which ultimately affects the sustainability of the service.

In social networks, the strict definition of churn may not apply, as users may or may not join another service after leaving the current one. Hence, the term *Continued Participation* is used to describe the future engagement of users within a social network. *Attrition* or *Defection* could also be used instead of churn, which indicates the gradual loss of customer or workforce in a certain system. But due to the enormous quantity of paradigms these terms are applied on, we inclined to use a term that uniquely describes the phenomenon in the online social networking paradigm.

Factors that influence participation continuation in social networks can vary from service to service. Graph based features can play a big role in predicting participation in those services which maintain an extensive architecture of relationships among the users like Facebook- whereas the frequency of activities plays a bigger role in the prediction task in services like forums and discussion boards. Demographic information, contents of texts, and timelines within user lifecycles can contribute significantly depending on which paradigm the prediction task is going to be performed.

In this work, we explore the difference between churn in traditional industries and continued participation in social networks. We focus on the special aspects of various social networks that can contribute to the participation prediction in different paradigms from previously conducted research works. We discuss the features and the techniques that have already been used to predict future participation along with the strengths and weaknesses of each approach. We also briefly discuss the challenges that are faced to predict future participation in social networks and possible research opportunities in this field.

Churn vs Participation Prediction

Churn has been one of the most researched topics in a wide range of industries like telecommunication sectors, banking, internet service providers etc. In these sectors, churn is generally referred to as customer loss. More specifically, in telecom sectors, churn represents the phenomenon of losing a company's customer to its competitor (Mozer et al., 1999; Hwang et al., 2004). To analyze the risk of churning, the term *Churn Rate* is defined, which is the number of customers who left the service divided by the number of total customers in that time period.

Churn analysis receives a high interest in these industries because it is believed that loss of customers is loss of revenue. Several studies have been conducted (Mozer et al., 1999; Coussement and den Poel, 2009; Buckinx and den Poel, 2005) and they suggest that:

- Retaining existing customers is much less expensive than winning new customers

- New customers tend to be less profitable than the existing customers

Which eventually suggests that a company should identify the subscribers who are going to churn and provide them with tailor-made incentives so that they do not leave and thus keep the expenses down.

In telecom industries, the term "Churn" is used extensively; because a regular subscriber is generally attached to one single carrier; that is, one subscriber does not carry two operators at once. This is not the case for online social networks, as a recent study from Pew Research Center found out that more than half (52%) of the internet users in 2014 subscribe to two or more social media sites (i.e. Facebook, Twitter, LinkedIn etc.) and this number is increasing every year (in 2013 42% of the internet users subscribed to two or more social media sites) (Duggan et al., 2015). Another study (Mander, 2015) states that a user actively subscribes to 2.8 social media sites while having 5.54 accounts on average in 2014. Hence, the general definition of churn does not apply to the cases of social networks and the term *Continued Participation* is devised. Although, churn is defined differently in different industries; for example, 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" (Herrera and Znati, 2007) which closely relates to the definition of continued participation in social networks. Still, due to the diversity of the definitions of churn over different industries, continued participation will be used interchangeably with churn further in this study.

Churn in Telecommunication Industries and Its Transition into Social Networking Paradigms

In the telecommunication industry, churn prediction is often conducted with a purely feature based approach (Burez and Van den Poel, 2009). These features are derived from factors in the service industry like pricing, convenience, customer service etc. along with customers' demographic factors e.g., age and gender (Keaveney, 1995).

A study on Korean mobile telephony market identified determinants that caused outbound churn from a cellular operator (Kim and Yoon, 2004). The initial hypothesis of this study was that a higher rate of churn signifies that the market is overall functioning well. It introduced a binomial logit model based on discreet choice theory to predict two aspects of churning- actual churn and loyalty which is described as the willingness to recommend current service. The study introduced two types of independent variables: service attributes i.e. call quality, tariff level, billing etc. and demographic variables i.e. age, sex, income, duration of subscription etc. These features were based on a study conducted by Gerpott, Rams and Schindler (Gerpott et al., 2001) which was done on the German cellular phone market and showed that customer retention, loyalty and satisfaction were causally interlinked. They also showed that service price, benefits and lack of portability also had significant effects on customer retention. The data Kim and Yoon used was a sample of 973 subscribers belonging to 5 mobile carriers in Korea. As the data was analyzed, the authors found a significant relationship between churn and service attributes- more precisely, three of them: call quality, handset type and brand image effect. Subscription duration is the only demographic variable that had a strong relationship with churn. They also found the same relationship between loyalty and the three service variables, but there was no significant contribution of subscription duration in a customer's loyalty value.

Another study conducted by Hadden et al. (Hadden et al., 2006) focused on using customer complaints data of a telecommunication network company to extract features and use machine learning in three different ways to predict churn in that network. The 24 independent variables authors used were not mentioned because of the data protection and the sensitivity of the data, but the authors loosely described the classes of those variables i.e. related to types of complaints, the number of days the complaint ran over the resolution date, the number of complaints made in a certain period of time etc.

As stated in the survey study conducted by (Karnstedt et al., 2010) social influence among individuals is neglected while studying churn in telecommunication industries which can be analyzed by examining the social networks of a user. In telecom industries, service providers often offer reduced call rates to the subscriber's friends and family, and even to the other subscribers of the network- which indicates that there is an innate social influence on whether a subscriber will churn from a certain network or not in the future. Studies have been conducted to introduce these community effects to replace or extend previously used feature based techniques (Kawale et al., 2009; Dasgupta et al., 2008), mostly by information diffusion models. These models predict the flow or propagation of information or discussion topic generated by outside world events or by the resonances in an existing social networking community by analyzing interpersonal communication among users (a broad summary of information diffusion models is presented by Gruhl et al. (Gruhl et al., 2004)). The idea behind these studies was that a subscriber is more likely to churn if he or she is connected to other users who have already churned from that network.

Continued Participation in Online Social Networks: Challenges and Opportunities

The loss of subscribers in an online social network is directly linked to the loss of social capital on that network. Constant et al. (1996) observed that users in some kind of social network invest their time sharing views or opinions or simply participating in a discourse without expecting any immediate return from the network. In sociology this type of activity is defined as the "Gift Economy" (Rheingold, 2000), which, in contrast to the the service or commodity economy, is not driven by exchanging service or commodities for monetary benefits, rather driven by the expectations of social contracts. Several motivations drive users to participate in this economy of gift transactions, for example, the expectations of future payback in terms of new information and social interaction, recognition as a source of valuable information from the peers or idea diffusion among other users in the community; and when these expectations are not met, users tend to leave the community, thus hurting the social capital of the network in the process. Social networking services lose revenue when users leave their network, just like other industries; but this loss of social capital poses a greater threat to the services as this threatens the survival of the social networks in the long run.

One of the challenges in predicting continued participation in online social networks is that there are no predefined “triggering events” (Gustafsson et al., 2005) in social networks as there are in telecom sectors. In telecommunication services, a subscriber is bound by a service contract or he buys credits before using the service. When the contract expires, or the credit dries up, churn is triggered based on the other factors like service quality, tariffs or poor customer experience. In social networks, users are weakly tied by a non-binding social contract (Constant et al., 1996). A user can leave a social network any time without incurring any kind of explicit monetary penalty, and can again join the network any time as there is low-entry barrier to join any social networks. This absence of triggering events makes it more difficult to predict continued participation in social networks than predicting churn in telecom or any other industries.

Another challenge while predicting continued participation in social networks is the diversity and the growth of the social networks (Karnstedt et al., 2011). There are chatrooms, discussion boards, community forums, photo and video sharing websites, blogs, massively multiplayer online games, online courses and many others which accumulate two or more of these services into them. The inner structures of these services are highly diverse and complicated. Discussion boards and blogs are mostly for sharing ideas and views by posts and replies in threads, and interpersonal communication among the users in these services are generally sparse- whereas chatrooms and online games depend mostly on the dense interpersonal communication among the users. Also, there are hierarchies of participation continuation in most of these services: a user can stop communicating with a single user or a set of users, or he can stop participating in a forum or a single thread, or he can leave the entire network.

One other challenge that makes continued participation prediction in social networks more difficult than predicting churn in the telecom sector is that in social networks, continued participation, as the name suggests, is a continuous process. A user does not suddenly drop off of a social network, it happens over a significant period of time. The amount of activity a user performs in a social network over a certain period of time is a good predictor of that user’s future participation (Sadeque et al., 2015). As there is no certain triggering event in social networks as there is in telecom services, participation discontinuation happens over a certain period of time in which the user gradually decreases his or her participation in the community and eventually stops participating at all.

Due to these challenges, continued participation prediction is still a largely unexplored attribute of social networks and thus gives us a massive research opportunity in this field. There have been some works to predict future participation in some popular paradigms like micro-blogging (e.g., Twitter; (Mahmud et al., 2014; Chen and Pirolli, 2012)) and massively multiplayer online role playing games (e.g., EverQuest II; (Kawale et al., 2009)), but some of the paradigms like health forums (i.e. DailyStrength) are still mostly unexplored. Quite a number of these services provide their data for commercially nonprofit and research purposes, and there is a huge opportunity to apply data mining and natural language processing in these data to establish a successful continued participation prediction models for these social networking paradigms.

Approaches to Predict Continued Participation in Social Networks

Several studies have been conducted to predict continued participation in different online social networking paradigms. Approaches in these studies differ based on the paradigms they were applied to. We divided these approaches in three major tracks:

- Linguistic and Behavioral Analysis
- Qualitative and Demographic Analysis
- Graphical Analysis of Activities and Interpersonal Relationships

Most of the research surveyed here apply one or more of these analysis approaches to predict continued participation.

Research	Task	Analysis	Model	Features	Paradigm
Joyce and Kraut, 2006	Discover significance of various characteristics of user's first post and the replies to that on future participation	Linguistic	Probit	Number of replies to the post, characteristics of the post and the replies like length in words, status i.e. question or testimonial, emotional tone etc.	UseNet News-groups
Arguello et al.. 2006	Discover influence of replies posted on user's initial post	Linguistic	Probit	Group identity, group size, volume, cross posting, user status, post characteristics i.e. rhetoric, question, testimonial, emotional tone etc.	UseNet news-groups
Mahmud et al., 2014	Predict social engagement behavior by means of response and retweet	Linguistic, Behavioral	Support vector machines	Psycholinguistic categories from LIWC of posted tweets	Twitter
Danescu-Niculescu-Mizil et al., 2013	Analyze user's linguistic change in a community and predict user lifecycle	Linguistic, Behavioral	Language models, logistic regression	Cross-entropy, Jaccard self-similarity, adoption of lexical innovation, first person singular pronoun, number of words	BeerAdvocate, RateBeer
Chen and Pirolli, 2012	Explore factors influencing engagement of Twitter users	Quantitative, Demographic	Regression	number of tweets, number of followers, number of followee, number of retweets, number of posted mentions, number of retweets and mentions from followers, the location etc at user level, and topic coherence	Twitter
Dror et al., 2012	Finding churners form a set of new users	Quantitative, Demographic	Random forest, logistic regression	Question features, Answer features, Gratification related features, demographic features	Yahoo! answers
Karnstedt et al., 2012	Predict discontinued participation of users based on the received value from community	Graphical, Interpersonal	J48 decision tree	In-degree, out-degree, closeness centrality, betweenness centrality, reciprocity, popularity, initialization, polarity etc.	borads.ie
Kawale et al., 2009	Predict churn from online games based on influence diffusion	Graphical, Influence	Modified Diffusion Model	in-degree, out-degree of nodes, weights of edges, influence vectors, spread factor etc.	EverQuest II
Sinha et al., 2014	Predict user attrition from massive online open courses using clickstream and forum activities	Graphical, Activities	Support vector machines	Number of nodes, edges, density, self loops, strongly connected components, central activity, in-degree centrality etc.	Coursera MOOC
Ngonmang et al., 2012	Predict churn using local community analysis	Graphical, Interpersonal	Support vector machines	degree of a node, local community size, average degree of nodes in local community, neighborhood size etc.	Skyrock.com
Sadeque et al., 2015	Predict continued participation in health forums	Temporal, Quantitative, Linguistic	Regression	Post count, reply count, time gap between activities, number of unigrams, questions etc.	DailyStrength
Hamilton et al., 2017	Analyze loyalty in online communities	Linguistic, Quantitative, Graphical	Random Forest	Post popularity, linguistic style, post contents etc.	Reddit
Liu et al., 2016	Churn prediction using relationship strength	Graphical, interpersonal	Clustering	# of posts, activity level, # of friends, proportion of churn friends, community structure etc.	Tencent Weibo
Milosevic et al., 2017	Churn prediction in mobile social games	Quantitative	Logistic regression, naive bayes, decision tree, random forest, gradient boosting	Activity features, monetization features, gameplay style features	Top Eleven- Be a Football Manager

Table 1: Summary of the reviewed research works on continued participation in social networks

Table 1 provides a summary of the research works that have been done in different social networking paradigms. The table provides a sneak-peek into the works with a list of the most significant features that have been used for the specific tasks. As can be seen from the table, most of the works are significantly different from each other in what to do, how to do it and where to apply. This raises a difficulty while comparing these works with one another. They also define inactivity or churn with drastically different time windows, i.e. Milosevic et al. (Milosevic et al., 2017) define churn of a user after only 14 days of inactivity, whereas Sadeque et al. (Sadeque et al., 2015) defines discontinued participation after a year of inactivity. All of these works along with their successes and limitations are explained in detail in the subsequent sections.

Linguistic and Behavioral Analysis

Linguistic features can contribute to a user's future participation in an online social network. The contents, emotional tone, length of the posts and replies a user has posted in a social network can be a good predictor of whether the user is going to leave the forum or not. Also, the responses the user receives from other users can play an important role in the prediction task. These features can be more prevalent in discussion based social networks like newsgroups or health forums.

Three major linguistic and behavioral (individual and/or community) attributes dictate participation prediction in social media: responses and their characteristics (Joyce and Kraut, 2006; Arguello et al., 2006; Danescu-Niculescu-Mizil et al., 2013), psycholinguistic features (Mahmud et al., 2014) and lexical innovation (Danescu-Niculescu-Mizil et al., 2013). Receiving replies or responses from peers has generally positive effects on continued participation; but if the replies are from newcomers or have too much complexity, the effect tends to be negative. On the other hand, replies with positive emotions have positive significance on future participation. Factors that affect response reception, i.e. group identity, volume, cross posting (simultaneously posting in multiple forums), user status (veteran or newcomer) and various post characteristics (length, complexity, usage of first and third person pronouns, being testimonial and/or topical) also serve as indirect attributes towards continued participation. Lexical innovation provides useful information on users' ability to adapt to an ever-changing social media paradigm, which can be used as a powerful predictor for future contribution.

Joyce and Kraut worked on predicting participation in Usenet newsgroups (Joyce and Kraut, 2006). In this work, they attempted to explain whether the responses received by the newcomers on their first posts in a newsgroup influence their continued participation in that newsgroup. They believed that a person's first participation in a newsgroup is the key to predict continued participation as the majority of the users in a newsgroup posts once in the user lifetime if they even post at all (Nonnecke and Preece, 2000). The authors worked on proving the six hypotheses that are derived from previous studies (Patterson, 1994; Davis and Holtgraves, 1983; Patterson, 1994; Davis and Holtgraves, 1983; Kiesler and Sproull, 1992) and tried to establish relationships between different independent variables obtained from the response of a user's initial post (i.e. its contents, quality, emotional tone) with the likelihood of the user posting again. They also hypothesized that characteristics of the initial post should also influence the continued participation as it explicitly influences whether it will get a reply and also the quality of the reply.

The authors collected six months of data from six online newsgroups. They identified users who posted first in their user lifecycle in the first three month of that six-month period, and then followed these users for the whole data collection period to identify whether they post again. They had 2777 records of new users. The only dependent variable used in the paper is Post-Again which denotes whether the user posted again or not. For independent variables, the authors used whether the post got any replies, and if yes, how many did it get. They also used the characteristics of the replies like length of message in words, whether the reply is a question or a statement, emotional tone using positive and negative emotion words from Linguistic Inquiry and Word Count (LIWC)[1] – the most commonly used lexicon and language analysis tool with 70 different word categories containing dozens to hundreds of words each – and agreeableness using assent and negation words. Furthermore, they used the characteristics of the initial posts (question, word count, emotional tone) as control variables.

[1] `http://www.liwc.net/`

Using a probit analysis they found out the group in which the initial message was posted influenced the likelihood of getting a reply, and that longer initial posts and receiving a reply both have positive significance over the prediction value. Effects of the characteristics of the replies also varied from group to group, but eventually the only characteristic that had some consistent significance over the prediction value in all groups was whether the reply was a question or not. The effect of emotional tones in both the initial post and the replies vary over the different groups- and thus the authors conclude that, out of the six hypotheses, only the first one ("Receiving a response to an initial post will increase the likelihood that the poster will post again") is supported, the second one ("An initial post that receives a response that provides information rather than asks a question will increase the likelihood that the poster will post again") was disconfirmed and the other four are not supported.

Arguello et al. attempted to find the factors that influence the number of replies a post gets in Usenet newsgroups which in essence captures the success of an online community (Arguello et al., 2006). As Joyce and Kraut suggests that the responses a user gets from his or her first post play a crucial role in his or her continued participation, this paper also tries to predict whether a user returns or not based on these analyses. The factors that the authors explored (and also used as features for learning) are divided into certain categories: Group-level factors (group identity, cross-posting and group size and volume), Individual level factor (Newcomer status) and message characteristics (Rhetoric, Topical coherence, linguistic complexity and word choice: Both linguistic complexity and word choice took advantage of the LIWC lexicon).

The data the authors collected are from 8 newsgroups that covers a wide range of topic and population from March 2001 to March 2002. They also collected structural data of the groups from Netscan project. In total the data had 6174 messages, each of them denotes the first post of a user that is not a reply to another message. The authors used the same probit analysis used by Joyce and Kraut to predict whether a message receives a reply, with four different sets of independent variables- where each set introduces a certain class of new features to the base model. The analysis showed that the group a user posts into has a significant influence on whether that user receives a reply, along with some characteristics of the post, i.e. being a testimonial increased the likelihood of getting replies by 10% and being a topical question increased the probability by 6%. Usage of longer and more complex sentences reduced the probability of getting a reply, whereas sentences containing more first person singular pronouns and third person pronouns increased the likelihood.

Arguello et al. included a new independent variable *gotReply* which denoted whether a user has received a reply or not in his or her first post. Based on this variable, they reported that getting a reply increased posters' probability of posting again by about 6.2%. They also found out that receiving replies from newcomers or having complex replies hurts the probability, whereas receiving replies with more positive emotion words actually improves the probability. Their concerns about the model were that the dataset was not large enough and the usage of bag of words as a measure of topical coherence, as this ignores syntax and context and only considers the usage of words.

The study conducted by Mahmud et al. attempted to analyze how word use can predict social engagement behavior in Twitter (Mahmud et al., 2014). The social engagement behavior is classified into two types: replies and retweets. The authors used the LIWC lexicon for the measure of word use in different psycholinguistic categories. They ran a Pearson Correlation Analysis between the LIWC categories and a user's past response rate and enlisted the categories that have the most statistically significant correlations. They also did the same thing between LIWC categories and the user's past retweet rate. For both of these analyses they found out some categories (anger, cognition, communication, anxiety, social process, positive feelings, positive emotions etc.) that have a noticeable statistical significance- both positive and negative- with the two variables. The authors also built a model to predict a user's future reply and retweet rate based on the usage of words of different psycholinguistic categories. They used correlation values of LIWC categories to past response rate and past retweet rate as independent variables and applied different regression models to predict the response rate and retweet rate. The authors introduced a binary classification model for engagement prediction with above-median response rate and retweet rate as 1 which denotes future engagement, and 0 as the opposite. They reported that the system can predict response and retweet rate with below 30% mean absolute error and can predict future engagement based on these rates with a 72-85% accuracy based on which LIWC categories were used.

Danescu-Niculescu-Mizil et al. proposed a framework that tracks linguistic change in social networks over time and how users react to these changes (Danescu-Niculescu-Mizil et al., 2013). For the framework they used a series of snapshot language models- one for each month of the user lifecycle in the community. These models are essentially bigram language models with Katz back-off smoothing(Katz, 1987) estimated from a held out subset of posts from each month. These models predict how surprising a user's language is compared to the general language style of a community using the cross-entropy of the post according to the snapshot language model:

$$H(p, SLM_{m(p)}) = -\frac{1}{N}\sum_{i} \log P_{SLM_{m(p)}}(b_i)$$

where $H(p, SLM_{m(p)})$ is the cross entropy, p is the post whose language is being scrutinized, $m(p)$ is the month when the post was written, $P_{SLM_{m(p)}}$ is the probability of the bigram b_i under the snapshot language model of that month $SLM_{m(p)}$ and $b_1, ..., b_N$ are the bigrams of p.

These snapshot models were applied to two social networking sites (BeerAdvocate and RateBeer[2]) and they revealed a two-stage lifecycle for the users: one is the learning phase where a user adopts the language of the community and the second one is the conservative phase where the user stops changing and lets the evolving norms pass by. The authors used this framework to detect the future participation of a user early in his or her career.

The change of language features over time provides a clear insight about why the usage of certain classes of words (first person pronouns etc.) can contribute to the prediction task of future participation. The study used this linguistic change as a predictor of user lifespan in social networks. The features they extracted are from the first w (w = a small number, i.e. 20) posts of a user, and the user is marked 'departed' if he or she abandoned the community before writing m more posts (m = another small number, i.e. 20). A user is marked as living if he or she stays long enough to post n posts- n is a relatively large number (i.e. 200). The study is not clear about how the authors marked the users who had written between 50 to 200 posts.

The features they used for learning are of several types, for example, there are five post-level features:

Cross-entropy Average cross-entropy of the post according to the snapshot language model of the month.

Jaccard self-similarity Jaccard self-similarity of the current post with past ten posts. This feature provides an insight of linguistic flexibility of the user. It is defined as

$$J(B_c, B_p) = \frac{|B_c \cap B_p|}{|B_c \cup B_p|}$$

where B_c is the set of bigrams of the current post and B_p is the set of bigrams of a previous post.

Adoption of lexical innovation takes value 1 if the post contains a lexical innovation in the community in the previous three months, and 0 otherwise. Innovation means usage of a bigram that has not been used in the community previously.

First-person singular pronoun this feature takes the value 1 if the post contains a first-person singular pronoun, or 0 otherwise.

Number of words number of words in the post.

For baselines they used two activity features: average time between posts and the month of the last review. Models used for predicting lifespan are binary logistic classifiers for both the baseline and proposed system. The system achieved better precision than the baseline for all the experiments they ran, yet the F1-measure is higher when they used post-level features along with the baseline as the recall gets significantly higher.

Hamilton et al. explored loyalty- which is a different take on continued participation- in online communities in Reddit[3] in their 2017 study (Hamilton et al., 2017). Two key aspects of loyalty, which they define as a combination of preference and commitment, are explored by the authors in this study: user loyalty, where a loyal user prefers a community over others, and community loyalty, where a

[2]publicly available data at http://snap.stanford.edu/data/

[3]http://www.reddit.com

loyal community retains its loyal users over time. User loyalty depends on individual user's linguistic and behavioral attributes. Upon analyzing the contents of the posts where loyal users post more than the vagrants (those who are not loyal to the community) do, the authors concluded that loyal users prefer posts with more esoteric contents- where the esotericity of a post is calculated by averaging the inverse document frequency of the noun phrases in the content. They also analyzed the stylistic difference of comments posted by different users- both loyal and vagrant, and came up with some interesting findings. They found out that in most of the subreddits, loyal users are more verbose, and use second person singular and first person plural pronouns, whereas vagrant users tend to use first person singular pronoun more. In their loyalty prediction task using only the first post of a user, the authors found out that these linguistic features are decent predictors of loyalty in 58% of the subreddits- which indicates that loyal users display affinities to certain stylistic elements really early in their user lifecycle. They also explored some quantitative and interpersonal attributes- we will discuss them later in this paper.

All of these studies suggest that linguistic and behavioral features obtained from analyzing a user's activities is a good predictor of user's future participation in online social networks. These studies have identified the features that have the most significant influence over future participation, and some have presented experimental setups that showed the success of these features in participation prediction tasks.

Quantitative and Demographic Analysis

A user's demographic information like age, sex, location etc. can be a predictor of his or her future participation in online social networks, though studies suggest that demographic information may not play a big part in the prediction task (Mahmud et al., 2014; Sadeque et al., 2015). Many activity-related attributes can also be quantified and used as features for participation prediction like number of posts, number of mentions (hashtag mentions in Twitter), number of response activities (retweets/forwards), number of replies, number of followers and followees etc. Some of these features have some decent significance over future participation. For example, more interactions with followers in social media suggests more engagement in the future; and users with a higher number of submissions with a higher amount of gratitudes (i.e. more upvotes or likes) are less likely to churn- whereas users who only participate in popular contents are more likely to churn.

Chen and Piroli conducted a study where they explored factors that influence the engagement of Twitter users in the Occupy wall Street movement (Chen and Pirolli, 2012). They started with four hypotheses:

- Followers who are more active in general are more engaged with the movement
- Followers who exhibits related topic interests are more engaged with the movement
- Followers who had more interaction with other followers before the movement are more engaged during the movement
- The geographic location of the followers affects their engagement level

As for dependent variables, the study uses the number of follower's retweets of OccupyWallSt and the number of tweets containing movement related hashtags by October 3rd. To test the hypotheses, they introduced other independent features: number of tweets of a user, the number of followers of the user, the number of followees of the user, the number of retweets of the user, the number of mentions the user has posted, number of retweets and mentions from other followers of the movement, the location of the user, and TopicInt - the dot product of two normalized TF-IDF vectors computed from the tweets of the user and a random sampling of roughly 500,000 tweets containing four hashtags #sep17, #ows, #occupywallst and #occupywallstreet for finding topic relevance. TF-IDF (Term Frequency-Inverse Document Frequency) of a term t in a document d from a set of documents D is defined as:

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

where

$$tf(t,d) = \text{number of times term } t \text{ occured in } d$$
$$idf(t,D) = \log \frac{|D|}{|d \in D : t \in d|}$$

TopicInt is then defined as:

$$TopicInt = \sum_{i=1}^{n} tfidf(t, d_{i1}, D_1) tfidf(t, d_{i2}, D_2)$$

where $t \in \{\#sep17, \#ows, \#occupywallst, \#occupywallstreet\}$, D_1 is the tweets posted by followers before July 14th, D_2 is the random sample of tweets, $d_{i1} \in D_1$, and $d_{i2} \in D_2$. The study found strong support for the third hypothesis, that more interaction before the movement led to more engagement during the movement by analyzing all the independent variables and their individual feature values to predict the two dependent variables.

Dror et al. attempted to detect answerers that are about to quit Yahoo! Answers and they specifically focused on the new users whose activity history is less than or equal to one week (Dror et al., 2012). The data they used had 20000 examples with 10944 churners and 9056 non-churners. The authors used 64 features divided into 4 categories: Question features (category, length in words and characters, total and average number of questions a user has answered that were deleted, average number of stars of the questions answered, etc; 33 total), Answer features (number of answers submitted, length of answer in characters and words, day of year, day of month, day of week, hour of day, etc.; 10 total), Gratification related features (total number of best answers by the user in first week, number of thumbs up and thumbs down in answers, response of the asker in characters and words, etc.; 18 total) and demographic features (gender, age, zip; 3 total). They used naïve Bayes, SVM, logistic regression, J48 decision tree, random forest and KNN (K=10). Out of all the techniques they used, random forest generated the best performance, but they preferred logistic regression as it was simpler to implement, faster and lagged only a little behind random forest. The authors also attempted to find out the most important features out of the 64 using information gain of each feature and concluded what is naturally expected of the features: i.e. whoever posts more is less likely to churn, the more gratitude a user receives the less likely he is to churn etc. The only surprising finding they reported was that users that are involved in more popular contents (contents that have greater amounts of participation; i.e. higher number of questions and answers) are more likely to churn- and this conclusion was obtained by analyzing the positive correlation of question related features with churning.

Milos Milosevic and his peers built a churn prediction mechanism for addressing the problem of discontinued particpation in mbile online games in their recent study (Milosevic et al., 2017). They randomly selected two million players from a popular online game named *Top Eleven - Be A Football Manager* and used their in-game activities to identify churners. They focused on early churn prediction, and hence only used day-one data for the prediction task. They divided the users in two major groups: churner and non-churner. Users who have not participated in the game in 14 consecutive days were marked as churners. The authors explored three major sets of features: Activity features i.e. session count (number of sessions a user has made), playtime (total number of hours the user has played the game) and click count (number of times the user has clicked in a particular screen), monetization features i.e. the amount of soft currencies (morale boosters, treatments, rests etc.) the user has used on his team, and gameplay features i.e. percentage of time the user has spent on different gameplay attributes (auction, training, transfer etc.). The authors used five different learning models: logistic regression, naive bayes, decision tree, random forest and gradient boosting. Out of these five, gradient boosting performed the best, with the highest F-1 score and largest area under the curve. The authors also implemented a personalized message based churn prevention system, which reduced churn upto 28%.

The Hamilton et al. study mentioned in the previous section explored some quantitative attributes behind loyalty (Hamilton et al., 2017). The authors found out that loyal users tend to comment more in low-scoring posts in all the communities that were being experimented on, and in 95% of them, posts with more comments overall are popular in vagrants as they comment more in those posts. The most interesting finding in this attribute set is that users who eventually become loyal to a community write their first comment to that community on less popular posts before any further activity, and that differentiates them from vagrants. The early loyalty prediction task supports this finding, as classifiers with only the post scores as features have strong performance in differentiating loyal user from vagrants.

Graphical Analysis of Activities and Interpersonal Relationships

The attribute that largely makes participation prediction in social networks different from churn prediction in the telecommunication sector is the interpersonal relationship among the users in social networks. All social networks have some forms of interpersonal relationships, though some are relatively sparse. Representation of these relationships as graphs can introduce features that can be used to predict future participation in online social networks. Three major types of graphs have been used for participation analysis: interpersonal relationship graphs (Ngonmang et al., 2012; Karnstedt et al., 2011), activity graphs (Sinha et al., 2014) and information diffusion graphs (Kawale et al., 2009). Generally, users are represented as nodes in those graphs, whereas edges represent either interpersonal relationship, simultaneous activities or influence diffusion among these users. These graphs provide researchers with some useful attributes for participation prediction i.e. degrees of nodes, various types of centrality, graph density, number of strongly connected components and self loops, positive and negative influence diffusion etc. In general cases, degree, initialization and popularity of a node in an interpersonal relationship graph are closely associated with non-churning of the corresponding user, whereas low number of nodes, edges, strongly connected components (SCC) and self loops along with low graph density in an activity graph contributes more to churning of the users participating in the activities. Diffusion models are useful in improving the performance of a participation prediction model, which is shown by (Kawale et al., 2009).

Karnstedt et al. (Karnstedt et al., 2011) explored the relation between how a user is valued in a community and the probability of a user stopping participation in future in a popular Irish forum site `http://boards.ie/`. The value of a user in a community is constituted from various user features. The authors used a week as the unit of user lifecycle, and the prediction of churn in this lifecycle is a continuous process. The authors defined churn (instead of discontinued participation) as the phenomenon of a user's activity dropping below a certain threshold:

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

Where $\mu_C(v_i)$ 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 $0 \leq T(S) \leq 1$.

Features used in this study are entirely graph-based: the authors introduced a connected weighted directed graph $G(V, E)$ where V is the set of vertices or nodes and E is the set of edges; and each node represents a user and each edge represents a reply directing to the node that represents the user who created the original post where the reply was posted. The features extracted from this graph are:

In-degree Number of incoming connections to a given user v_i

Out-degree Number of incoming connections to a given user v_i

Closeness Centrality The importance of the user based on their location in the reply graph. 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

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

Closeness centrality is the inverse of l_i

$$C_i = \frac{1}{l_i}$$

Betweenness Centrality Measure of a user being a conduit or a broker 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 $v_i \neq v_y$. Betweenness for v_i is defined as:

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

Reciprocity Average time it takes for a post from a user to be replied to

Average post in initiations Average length of conversation in words on threads initialized by a user

Average post in participations Average length of conversation in words on threads a user takes part in

Popularity Percentage of the posts the user got replies to. Popularity of user *i* over time period $[t_1, t_2]$ is defined as:

$$pop_i(t_1, t_2) = \frac{|r(pst_x, pst_y, t_{xy})|}{|Pst_i(t_1, t_2)|}$$

where $Pst_i(t_1, t_2)$ denotes the set of posts user *i* has posted over the time period $[t_1, t_2]$, $r(pst_x, pst_y, t_{xy})$ denotes $pst_y \in Pst(t_1, t_2)$ is a reply of post $pst_x \in Pst_i(t_1, t_2)$ and there was a delay of t_{xy} time units (minutes) between them.

Initialization Popularity of threads initialized by a user. It is defined as

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

where $|thr_l|$ is the length of thread *l* in words and $|init_i(t_1, t_2)|$ is the length of the set of the threads initialized by user v_i over time period $[t_1, t_2]$

Polarity Average sentiment of a user's post over the past 6 months. The authors used Sentiwordnet's sentiment lexicon[4] for sentiment measurement and then took the average of the polarity measure of each post in the collection. Mathematically,

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

where *c* is the number of unique terms in post p, $pos(T)$ is the positive weight of the term T from Sentiwordnet and $neg(T)$ is the negative weight.

The study used J48 tree to predict churn which is a binary dependent variable obtained from the churn probability calculated from the features and then comparing to a certain threshold σ; if the churn probability exceeds the threshold, the user is classified as a churner. The authors used three thresholds- 0.2, 0.5 and 0.7, and showed that performance increases as the threshold rises. The authors also found out that in-degree and out-degree features are the most important, and initialization also has significant contribution. They also concluded that a consistent level of popularity is closely associated with a user not churning from the online social network. They used these features and techniques to predict per forum churn in different types of forums and found out that the effect of the threshold value contributes differently based on the type of the forum. The authors also analyzed neighborhood churn for different kinds of forums, where they tried to find a correlation between a user's churn probability and his or her neighborhood's churn probability. They concluded that the more active the forum is, the less correlation between user churn and neighborhood churn exists.

Kawale et al.(Kawale et al., 2009) explored the phenomenon of player churn in a Massively Multiplayer Online Role Playing Game called EverQuest II. The authors proposed a churn prediction model that uses social influence among players and their personal engagement in the game. They built a weighted graph $G(V, E)$ with vertices V and edges E based on the activities of the players in the month of August in 2006. Each node $v \in V$ in that graph represents a player, and an edge $e \in E$ between two nodes means that these two players has taken part in a quest together. The weight of an edge represents the points shared by the two players- which in a sense represents the strength of tie between them. They hypothesize that player engagement of churners decreases over time until they finally churn, and churn propensity of a player increases with the increasing number of churner neighbors: both of these hypotheses were proven by the analysis of the dataset they used.

The study introduces a Modified Diffusion Model (MDM) on the graph to formalize the influence that a player has over his neighbors and how the influence propagates over the graph. Each node on the graph has an influence vector (with direction between nodes) and has a positive (pi) or negative (ni) component (denotes positive and negative influence) based on whether the influencing user is a churner or a nonchurner and represents how much a user is influenced for or against the game. The spread factor γ represents the portion of influence that is transferred from a user to the network. The total influence of the graph remains unchanged over propagation of the model, only the positive and negative influence values change. Churners are initialized with $ni = 1$ and $pi = 0$ and nonchurners are initialized with $ni = p$ and $pi = 1 - p$. After that, the authors used their MDM

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

to diffuse the influence using the following algorithm:

---

if $S_x(t) < 0$ **then**
  $i = min(ni(x), pi(x))$
  $pi(x) = pi(x) - i * (1 - \gamma)$
  for $y \in N(x)$ **do**
    $$ni(y) = ni(y) + i * \gamma * \frac{e_{xy}}{e_x}$$
  end for
end if
{if $S_x(t) > 0$, interchange $ni(.)$ with $pi(.)$ in above step}
{if $S_x(t) = 0$, do not consider the node}

Here $N(x)$ is the number of edges from x, e_{xy} is the weight of the edge between x and y, e_x is the sum of weights of all edges from x and $S_x(t) = \frac{d\Phi_x(t)}{dt}$, the slope of $\Phi_x(t)$, x^{th} player's engagement in time t.

The prediction task was done using three different models: Simple Diffusion Model- where there is only one influence or energy *e* which is initialized with 1 for churners and 0 for nonchurners and with a γ of 0.7 and the propagation follows the same algorithm as MDM but does not consider the weights of the edges, Classification based on Network and Player Engagement- where there is no propagation of influence at all and prediction is done using the raw network features like in-degree, out-degree of nodes, weights of the edges etc. and their Modified Diffusion Model. They concluded that among the three systems they experimented, their MDM system outperformed the other two as it combines both social influence and player engagement (Simple diffusion model uses only social influence whereas Network and player engagement model uses only engagement features) and provided a significant improvement in prediction accuracy.

The study conducted by Sinha et al. attempted to leverage combined representations of video clickstream interaction and forum activities to fundamentally understand the traits that are predictive of decreasing engagement over time (Sinha et al., 2014). The predictive model was built using one Coursera[5] Massive Open Online Course (MOOC) as training data and it was tested on five other Coursera MOOCs. The variables used to capture behaviors were of two kinds: Clickstream activities (Play, Pause, SeekForward, SeekBackward etc.) and Discussion forum activities (Post, Comment, Upvote, Downvote etc.). All the MOOC events were sorted using their timestamp to gain a simple sequentially ordered time series to analyze behavioral pattern of the students. They also extracted N-grams from the interaction footprint sequence- ('n' consecutively occurring MOOC activities where n = 2 to 5), proportion of video viewing activities and proportion of discussion forum activities among all interactions in each of the participating week of each student.

Using the clickstream activities, the authors constructed a graph with the activities as nodes in that graph and two consecutive activities having an edge between them. They extracted as graph based features the number of nodes $|V|$, edges $|E|$, density $\frac{|E|}{|V|(|V|-1)}$, number of self loops, number of strongly connected components (SCC), central activity (top three activities by in-degree centrality, the fraction of other nodes that a node's incoming edges are connected to) and central transition (edges with maximum betweenness centrality, the sum of all-pair shortest paths that pass through an edge). They also created two experimental setups: one uses data from the current week (Curr) and the other uses data from the beginning till the current week (TCurr). Analyzing the features, the authors conclude that dropout is higher for students having low number of nodes, edges, SCC and self loops, low activity graph density, low proportion of active forum and video viewing activity. They also mentioned that student dropout is higher if they join in later course weeks and have a sparse activity graph. For the prediction task, they used two models, one using the baseline Ngrams, and other using the graph features and as for learning algorithm they used Lib-SVM with radial basis kernel function. From the results collected from these models, the authors concluded that the graph model performs significantly better than the Ngram one, and students' attrition is more strongly influenced by the most recent week's behavioral pattern.

Ngonmang et al. presented an algorithm that accurately detects overlapping local communities in social graphs to predict the churn in an online social network (Ngonmang et al., 2012). They used local community analysis (a local community of a node n in a graph G is a subset of nodes that starts at n_0 and that has a high internal density (δ_{in}, the number of internal links of a sub-graph divided by the number of vertices) and low outgoing density (δ_{out}, the number of outgoing links divided by the

[5] `http://coursera.org`

number of nodes)) to maximize the quality function of their prediction algorithm. They used a popular French social networking service called Skyrock[6] for the source of their data and the graph they built out of that data had a massive 31.3×10^6 nodes (representation of a user) and 1.17×10^9 friendship links among them. They filtered the graph and computed the local communities from the graph and randomly selected 50000 users to create the learning dataset where each user (or node) had some common attributes like degree, community size, proportion of active nodes in the community etc. They used an overall 21 features (degree of a node, local community size, average degree of nodes in local community, neighborhood size, etc.) for the learning model and used libSVM with a radial basis function kernel. The authors divided the features into multiple feature sets, used those sets to compare among themselves and finally used all of the features to produce the best performance measures. Their results suggest that although local community features like proportion of inactive users in the community have some effects, most of the performance boost comes from the single-node features like degree, clustering coefficient, neighborhood size etc.

A recent study by Liu et al. (Liu et al., 2016) introduced a Churn Impact Iterative Model (CIM) to predict churn in social media using relationship strength. They attempted to focus on the influence of friends of a user on his or her churning from social media. This study was targeted for establishing community influence as a predictor for churning beyond the person-level attributes. The study focuses on some of the most basic attributes of a friendship network, and found correlation between them and churning probability. For example, the more friends a user has, the less likely he or she is to churn from that community; and the higher the churn vs. non-churn ratio of friends in a user's network, the more likely the user is to churn from that community (the exception to this case is that if the ratio is zero, then the user is more likely to churn as this is an indication that the user has no friends). Their analysis also showed that churn probability of a user in an overall inactive community is significantly higher than that of a user from an active community.

Last but not the least, we go back to the study that we mentioned in previous sections: Loyalty in online communities by Hamilton et al. (Hamilton et al., 2017). The authors explored interaction networks among users in certain communities to analyze community level loyalty. One of the basic attributes of basic communities is that they tend to be smaller than a non-loyal community- the median loyal community is 39% smaller than the median non-loyal community. Also, loyal communities are more active- per-user post and comment count is higher in loyal communities than in non-loyal communities. Conversations in loyal communities are also longer, but with fewer contributors.

The authors created user-user interaction networks in both loyal and non-loyal communities. Their analysis of these networks showed that networks in loyal communities are denser and have less local clustering- indicating that loyal communities are more tight-knit and cohesive. The edge density of interaction networks in loyal communities is significantly higher than that of the non-loyal ones- which indicates that the average user in a loyal community interacts with more users. This interaction is skewed by highly active users though, as these active users communicate with more users on average than the active users in non-loyal communities do.

Studies conducted using graphical analysis of interpersonal relationships have suggested that interpersonal relationships have a huge significance over a user's participation in a social network. Exploiting these interpersonal features has a drawback- it can not be used in social networks where these kinds of relationships do not exist. Graphical analysis of activities, on the other hand, can be implemented in a multitude of paradigms.

Limitations of Previous Works

Most of the works performed in the field of participation prediction suffer from a handful of limitations. Demographic information obtained from users' social networking accounts or in any other way is a poor predictor of future participation as it is already proven in several studies- yet it has been used as a feature for the classification models. Also, most of these works do not have sufficiently large data sets. Not a single finding of a particular research work has been verified in a different paradigm, or in a different social networking website of the same paradigm. Hence, adaptability of all these methods can be questioned.

There is one dimension that can offer new op-

[6] http://skyrock.com/

portunities in this field, yet is often ignored or unexplored- time. Temporal analysis of a user's lifecycle can be a massive predictor of user's future participation in a social network. Some of the previous works has done some timeline analysis i.e. Danescu-Niculescu-Mizil et al. created snapshot language models based on time to analyze the change of user's language over his or her lifecycle and used it as a predictor of future participation (summarized in section *Linguistic and Behavioral Analysis*) (Danescu-Niculescu-Mizil et al., 2013). Kawale et al. used the timeline of the activities of a gamer participating in massively multiplaying online role playing games as an important predictor of the gamer's future participation in that game (Kawale et al., 2009). Sadeque et al. introduced a prediction model that implemented linguistic, quantitative and temporal features obtained by observing a user for a certain period of time (observation period) to predict continued participation in the popular support group based online health forum DailyStrength[7] and found out that time related features like the number of days berween the point at which the user created his or her account and his or her first post and the time between a user's last post and the end of the observation period has a much more significant impact than the other features and the impact is consistent over different observation period (Sadeque et al., 2015). Lampe and Johnson introduced the time gap between two activities of a user as an indicator of socialization in their study (Lampe and Johnston, 2005). The work of Sinha et al. (Sinha et al., 2014) also tried to incorporate timeline is their system- their features were based on timestamped order of activities, but it did not consider the timegap between two consecutive activities- as the effect of a previous task over the next task fades out as the timegap between these two increases.

Conclusion

Churn or continued participation has been one of the most researched topics in service industries for quite some time, where losing customers or subscribers results in loss of revenue. It is not as widely researched in online social networking services as in telecom or some other industries due to the variety of challenges the paradigms presents, but some work has been done over the years, and it is becoming an emerging field in the study of social networking as a whole. In this study we attempted to present a broad picture of research in participation prediction in different industries along with the research opportunities social networks present in this field. We presented the importance of predicting continued participation in social networks as it is directly related to the health of the social networks. We also differentiated between churn in telecom industries and continued participation in social networks and highlighted the challenges researchers may face in this field of study. We summed up the most important research works on continued participation prediction in social networks and identified the approaches that has been followed in these works. This study will help future researchers to get a concise view into the current situation of the field and encourage current researchers to explore different approaches and get the best out of them.

[7] `http://dailystrength.org`

References

Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. 2006. Talk to me: Foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 959–968, New York, NY, USA. ACM.

Wouter Buckinx and Dirk Van den Poel. 2005. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual {FMCG} retail setting. *European Journal of Operational Research*, 164(1):252 – 268.

Jonathan Burez and Dirk Van den Poel. 2009. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3):4626–4636.

Jilin Chen and Peter Pirolli. 2012. Why you are more engaged: Factors influencing twitter engagement in occupy wall street. In *Sixth International AAAI Conference on Weblogs and Social Media*.

David Constant, Lee Sproull, and Sara Kiesler. 1996. The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organization science*, 7(2):119–135.

Kristof Coussement and Dirk Van den Poel. 2009. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*, 36(3, Part 2):6127 – 6134.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 307–318, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit A Nanavati, and Anupam Joshi. 2008. Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, pages 668–677. ACM.

Deborah Davis and Thomas Holtgraves. 1983. Perceptions of unresponsive others: Attributions, attraction, understandability, and memory of their utterances.

Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. 2012. Churn prediction in new users of yahoo! answers. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 829–834. ACM.

Maeve Duggan, Nicole B. Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. 2015. Social media update 2014: While facebook remains the most popular site, other platforms see higher rates of growth. `http://www.pewinternet.org/files/2015/01/PI_SocialMediaUpdate20144.pdf`. Accessed: 2015-09-26.

Torsten J Gerpott, Wolfgang Rams, and Andreas Schindler. 2001. Customer retention, loyalty, and satisfaction in the german mobile cellular telecommunications market. *Telecommunications Policy*, 25(4):249 – 269.

Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. 2004. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 491–501, New York, NY, USA. ACM.

Anders Gustafsson, Michael D Johnson, and Inger Roos. 2005. The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. *Journal of Marketing*, pages 210–218.

John Hadden, Ashutosh Tiwari, Rajkumar Roy, and Dymtr Ruta. 2006. Churn prediction using complaints data.

John Hadden, Ashutosh Tiwari, Rajkumar Roy, and Dymitr Ruta. 2007. Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, 34(10):2902–2917.

William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. *arXiv preprint arXiv:1703.03386*.

O. Herrera and T. Znati. 2007. Modeling churn in p2p networks. In *Simulation Symposium, 2007. ANSS '07. 40th Annual*, pages 33–40, March.

H. Hwang, T. Jung, and E. Suh. 2004. An ltv model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert Systems with Applications*, 26(2).

Elisabeth Joyce and Robert E. Kraut. 2006. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3):723–747.

M. Karnstedt, T. Hennessy, J. Chan, and C. Hayes. 2010. Churn in social networks: A discussion boards case study. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 233–240, Aug.

Marcel Karnstedt, Matthew Rowe, Jeffrey Chan, Harith Alani, and Conor Hayes. 2011. The effect of user features on churn in social networks. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 23:1–23:8, New York, NY, USA. ACM.

S. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(3):400–401, Mar.

J. Kawale, A. Pal, and J. Srivastava. 2009. Churn prediction in mmorpgs: A social influence based approach. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*, volume 4, pages 423–428, Aug.

Susan M. Keaveney. 1995. Customer switching behavior in service industries: An exploratory study. *Journal of Marketing*, 59(2):pp. 71–82.

Sara Kiesler and Lee Sproull. 1992. Group decision making and communication technology. *Organizational behavior and human decision processes*, 52(1):96–123.

Hee-Su Kim and Choong-Han Yoon. 2004. Determinants of subscriber churn and customer loyalty in the korean mobile telephony market. *Telecommunications Policy*, 28(910):751 – 765.

Cliff Lampe and Erik Johnston. 2005. Follow the (slash) dot: Effects of feedback on new members in an online community. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '05, pages 11–20, New York, NY, USA. ACM.

Ya Liu, Yan-Bing Liu, and Yun-Peng Xiao. 2016. A churn prediction iterative model in social network based on relationship strength. *Joint International Conference on Artificial Intelligence and Computer Engineering*.

Jalal Mahmud, Jilin Chen, and Jeffrey Nichols. 2014. Why are you more engaged? predicting social engagement from word use. *CoRR*, abs/1402.6690.

Jason Mander. 2015. Gwi social summary: Globalwebindexs quarterly report on the latest trends in social networking. `http://tinyurl.com/nd9ur8r`. Accessed: 2015-09-26.

Milo Milosevic, Nenad Zivic, and Igor Andjelkovic. 2017. Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications*, 83:326 – 332.

Michael Mozer, Richard H Wolniewicz, David B Grimes, Eric Johnson, and Howard Kaushansky. 1999. Churn reduction in the wireless industry. In *NIPS*, pages 935–941.

Blaise Ngonmang, Emmanuel Viennet, and Maurice Tchuente. 2012. Churn prediction in a real online social network using local community analysis. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 282–288, Washington, DC, USA. IEEE Computer Society.

Blair Nonnecke and Jenny Preece. 2000. Lurker demographics: Counting the silent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 73–80, New York, NY, USA. ACM.

Miles L. Patterson. 1994. Interaction behavior and person perception: An integrative approach. *Small Group Research*, 25(2):172–188.

Howard Rheingold. 2000. The virtual community: Homesteading on the electronic frontier.

Farig Sadeque, Thamar Solorio, Ted Pedersen, Prasha Shrestha, and Steven Bethard. 2015. Predicting continued participation in online health forums. In *SIXTH INTERNATIONAL WORKSHOP ON HEALTH TEXT MINING AND INFORMATION ANALYSIS (LOUHI)*, page 12.

Tanmay Sinha, Nan Li, Patrick Jermann, and Pierre Dillenbourg. 2014. Capturing attrition intensifying structural traits from didactic interaction sequences of mooc learners. In *Proceedings of the 2014 Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses*.