Título do trabalho a ser apresentado à CPG para a dissertação/tese

Nome completo do Autor

Dissertação/Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre/Doutor em Ciências

Programa: Nome do Programa

Orientador: Prof. Dr. Nome do Orientador

Coorientador: Prof. Dr. Nome do Coorientador

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES/CNPq/FAPESP

São Paulo, fevereiro de 2011

Título do trabalho a ser apresentado à CPG para a dissertação/tese

Esta é a versão original da dissertação/tese elaborada pelo candidato (Nome Completo do Aluno), tal como submetida à Comissão Julgadora.

Título do trabalho a ser apresentado à CPG para a dissertação/tese

Esta versão da dissertação/tese contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 14/12/2010. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Profª. Drª. Nome Completo (orientadora) - IME-USP [sem ponto final]
- Prof. Dr. Nome Completo - IME-USP [sem ponto final]
- Prof. Dr. Nome Completo - IMPA [sem ponto final]

Agradecimentos

Texto texto. Texto opcional.

Resumo

SOBRENOME, A. B. C. **Título do trabalho em português**. 2010. 120 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2010.

Elemento obrigatório, constituído de uma sequência de frases concisas e objetivas, em forma de texto. Deve apresentar os objetivos, métodos empregados, resultados e conclusões. O resumo deve ser redigido em parágrafo único, conter no máximo 500 palavras e ser seguido dos termos representativos do conteúdo do trabalho (palavras-chave). Texto texto. Texto texto.

Palavras-chave: palavra-chave1, palavra-chave2, palavra-chave3.

Abstract

SOBRENOME, A. B. C. **Título do trabalho em inglês**. 2010. 120 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2010.

Elemento obrigatório, elaborado com as mesmas características do resumo em língua portuguesa. De acordo com o Regimento da Pós-Graduação da USP (Artigo 99), deve ser redigido em inglês para fins de divulgação. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text.

Keywords: keyword1, keyword2, keyword3.

Contents

Lista de Abreviaturas

CFT	Transformada contínua de Fourier (*Continuous Fourier Transform*)
DFT	Transformada discreta de Fourier (*Discrete Fourier Transform*)
EIIP	Potencial de interação elétron-íon (*Electron-Ion Interaction Potentials*)
STFT	Tranformada de Fourier de tempo reduzido (*Short-Time Fourier Transform*)

Lista de Símbolos

ω	Frequência angular
ψ	Função de análise *wavelet*
Ψ	Transformada de Fourier de ψ

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction esciece15

Studies on social networks often use actions people take on other people's online content as evidence of social interactions for developing their models. In domains including Usenet [JK06], Wikipedia [BWCD11], and Facebook [GK09], explicit replies are interpreted as evidence of interpersonal interaction and social ties. These explicit reactions are also used in studies of influence online, such as predicting when an item is likely to be forwarded in Twitter (e.g., [SHPC10, CC12]).

Not all responses, however, are explicitly marked by the system. For instance, a post that is explicitly threaded as a reply to a particular post in a discussion forum might nevertheless address another post or posts. In Twitter, the primary focus of this paper, there are buttons for replying to and retweeting another user's tweet—but users might compose a new tweet that references another recently seen without hitting the reply button. Users might do this for a variety of reasons, from being inspired to write their own post on a topic they see coming up in their feed to using the system in ways not intended by the designer (such as copying and pasting content into a new tweet rather than pressing a retweet button).

Being able to identify these non-obvious, indirect responses might allow researchers to have a more accurate view of social interaction than explicit mechanisms provide. This might also improve overall estimates of users' responsiveness to others, for instance, at the individual level, they might indicate how desirable a user is as a follower: people might wish to have followers who are more likely to redistribute their content. Aggregating responsiveness of a user's followers at the ego network level could support better estimates of an individual's potential reach or influence [DR01] based on the responsiveness of their followers. Better responsiveness measures could also improve transmission probabilities in epidemiology-inspired models of diffusion in social networks [BRMA12].

This paper assesses the prevalence of non-explicit responses in a dataset drawn from Twitter, using a measure of normalized textual similarity between a user's tweets and recent friends' tweets based on tf-idf scores. Comparing this to the explicit responses provided by the system shows that explicit indicators of response (replies and retweets) in Twitter are in fact associated with high normalized similarity scores. Choosing conservative score cutoffs for predicting that a tweet is a response and manually inspecting high-scoring tweets that are not marked as responses suggests that explicit indicators miss at least 11% of reactions. Further, this varies between users: some users systematically fail to use formal response mechanisms, meaning that these users are under-represented in studies that rely on explicit indicators of response and under-counted when considering their potential as information spreaders. These results show that the problem of non-explicit responses is an important one with practical implications for understanding interaction and influence online.

1.2 Introduction www16

Understanding the evolution of users in a social network is essential for a variety of tasks: monitoring community health, predicting individual user trajectories, and supporting effective recommendations, among others. Many works aim at explaining these temporal aspects of evolution. Some adopt a point of view of the whole network and try to understand more general patterns of behavior [ZKK14, KGM10], while others adopt a more user-centric point of view and try to model [CHdZ10, PCL+07, PHT09, WCK+11] or predict [DNMWJP13] individuals' behavior.

These approaches often combine all available data into aggregate analyses of the whole community over its entire history. This can be a natural response to limitations in the amount of available data: many datasets capture a small part of the community's history [APG12]; timestamps may not be available [PCL+07, PES10]; snapshots may provide limited views of the community [CHK10]; or the community itself might be small [LKC08]. Aggregate time-based analyses are also a natural first way to address questions of community evolution.

However, we argue it is likely that many of these aggregated views are misleading. The conditions under which users join the community may vary greatly over time, and this might impact their behavior [MCT15]. Among other things, the popularity, purpose, features, interface, and algorithms can change: Wikipedia circa 2005 and circa 2015 are very different, as are Facebook of 2005 and 2015. Analysis—including some of our own past work—that fail to account for this change may miss important details of what is really going on.

We support this argument through an analysis of user effort in Reddit, one of the most popular and long-running online communities, based on a very large, recently released dataset of posting behavior. We address a number of questions commonly raised about users' effort in online communities: how active are users, how hard do they work, and what kinds of things do they do? In each case, we compare aggregate analyses of posting behavior to ones that treat users in Reddit as yearly cohorts, and views that focus on calendar time versus user-referential views that normalize behavior based on the creation date of a user. We also look at differences within yearly cohorts, seeking differences between shorter and longer-lived users.

We find that even simple accountings for time reveal additional insights about Reddit beyond what commonly performed aggregate analyses can provide. Users who join Reddit earlier post more and longer comments than those who join later, while users who survive longer start out both more active and more likely to comment than submit compared to users who leave Reddit early; none of these findings are obvious from commonly used analysis of user behavior.

Further, we find that aggregate analysis can be downright misleading. For instance, although average comment length decreases over time in an aggregate view, the comment length for surviving users increases over time in every cohort. Likewise, an aggregate analysis suggests that longer-lived users post more over time; this is not the case. Instead, users come into Reddit as active as they will ever be (somewhat akin to Panciera et al.'s finding that Wikipedians are "born, not made" [PHT09]), and the rise in average activity for surviving users over time is driven fully by lower-activity users leaving early.

We see this paper as both making specific contributions to understanding behavior in Reddit and a more general contribution around the importance of considering change over time in analyzing online communities.

Chapter 2

Related Work

2.1 Related Work esciece15

This question of identifying when a user is reacting to some other users' content can be considered a dual question to "is the user going to react to this message", a question often asked in studies around influence online. The usual approach to the latter question is to identify relevant message or network features in the set of (message, reaction), where the reactions are those tagged by the system (e.g., explicit retweets and/or replies in Twitter). Using these features, it is possible to build models that predict the likelihood of a reaction given a message.

Such studies often focus on computational models for predicting retweet behavior. For instance, Suh et al. [SHPC10] apply Principal Component Analysis to decompose tweets into a space of characteristics, showing that URLs, hashtags, the number of followers and followees, and the age of the account are correlated with retweet behavior. Comarela et al. [CC12] also find that previous responses to the same tweeter, the tweeter's sending rate, and the age of a tweet influence retweeting, proposing two ranking methods for reordering tweets to increase retweeting. Petrovic et al. [POL11] built a *passive-aggressive* classifier for answering that took into consideration social characteristics of the tweets' author as well as tweets' textual features, finding that social features are more informative. Peng et al. [PZP11] used *Conditional Random Fields* to model the probability of how a user retweets a message.

Other studies look at variations of the problem. Artzi et al. [APG12] applied *Multiple Additive Regression-Trees* and *Maximum Entropy Classifiers* to predict both retweets and replies, while Hong et al. [HDD11] model both the binary question of whether a tweet would be retweeted and the eventual number of retweets a message might accrue. Luo et al. [LOTW13] and Wang et al. [WLZL12] approach a similar problem: given a user and their followers, who will retweet a message generated by the user? Both created classifiers to predict the followers that would retweet a message. Liu et al. [LJ13] studied the social network of questions and answers in *Sina Weibo* looking for characteristics that are associated with a higher number of answers.

These prior works identify a number of useful features that researchers often take into consideration when developing their models. These include textual features of Tweets, user preferences or characteristics, and features of users' networks including pairwise relationships and graph structure. Table 2.1 presents a number of these features and the papers that have used them in response prediction. This paper's focus on the prevalence of implicit responses complements these works by identifying tweets that, although not marked as a response, are in fact likely to be real responses. Such tweets would appear as errors or noise to these models; methods for identifying them might improve both these models and our understanding of why these features matter. For instance, account age might turn out to predict retweet behavior mostly because more experienced users are simply more likely to press the retweet button than new users, rather than having a higher innate propensity to retweet.

When trying to identify non-explicit responses, having a model that explains which messages a user is most likely to be interested can be valuable; that is, the problem of understanding these

(message, user) relationships is related to the problem of understanding the (message, reaction) relationships. The main stream of research related to modeling user interests in Twitter is the feed personalization problem, defined by Berkovsky et al. [BF15] as creating mechanisms that promote and optimize exhibition of interesting content (messages or people, for instance) according to each user's particular preferences and context. In their survey, they break approaches to feed personalization into three main groups: approaches that consider the pairwise relationship between author and consumer of content, approaches that take into consideration the graph structure of the social network, and approaches that deal with textual information from the users.

As with studies of retweet prediction, feed personalization approaches often use indicators of tie strength as proxies for potential interest. Schaal et al. [SOS12] measure pairwise user similarity through tf vectors and topic similarity using LDA. Goyal et al. [GBL10] estimate pairwise influence probability based on the user activity (action log). There are a wide variety of such features; Gilbert and Karahalios [GK09] estimate pairwise tie strength based on Facebook data based on over 70 features in categories including intensity, intimacy, duration, reciprocal services, structural, emotional support, and social distance.

Network structure also plays an important role in feed personalization. Uysal et al. [UC11] developed a personalized tweet ranking method based on a retweet metric, useful in reordering feeds or distributing items to users more likely to retweet. Paek et al. [PGC+10] asked Facebook users about the perceived importance of items in their timeline, developed classifiers to identify important messages and friends, and studied the predictive power of a number of features including likes, number of comments, presence of links and images, textual information, and shared background information. Both the tie strength and network structure approaches rely on explicit interaction as a tool for estimating tie strength; just as with retweet prediction, being able to identify non-explicit responses might improve these models.

Most related to this paper are text-focused approaches. Text is commonly used in feed personalization, by comparing content similarity of Tweets or users to a user's previous activity. Hannon et al. [HMS11] developed a system for follower recommendation on Twitter based on tf-idf similarity between the users' newsfeeds. Burgess et al. [BMAC13] propose a system to automatically select users when creating lists. The method adopts tf-idf to compare content users generated, among other measures and evaluates the performance comparing user-made lists with those generated by the system. This work informs ours by providing evidence that tf-idf-based methods are useful in understanding attention and interest.

2.2 Time matters www16

2.2.1 Why accounting for time is important

Communities grow and, with time, die. For any community, its users play a role in its evolution, but they are also simultaneously affected by the evolution of the community. Untangling this interplay can help make sense of patterns of activity in a community.

One useful way to understand the evolution of a community and its users is through time, as it provides a linear account of the growth (or decay) of overall activity, types of content, social norms, and structure of communities. To account for time, users on online communities are differentiated based on their age, such as when modeling their preferences [ML13] or analyzing the evolution of their language [DNMWJP13]. These analyses uncover insights about the lifecycle of a user in a community: users' preferences and behavior change with their age in a community [PPET10], and their early experiences and activity shape future outcomes predictably [TL15, YC09, PHT09, MCT15].

However, much of past work on online communities ignores the time at which a user joins the community and analyzes all users together, irrespective of when they joined a community. This might be a mistake: communities may grow denser or sparser with time [LKF05], develop new norms [KGM10] and/or enact policies and rules guiding people's behavior [BJP08]. These changes mean that people experience different versions of a community at different times, which can, in

Table 2.1: *Some characteristics from online social networks that are commonly used to model users' behavior.*

Characteristic	Description
URL	Presence of a link in a tweet. [APG12, CC12, PZP11, POL11, SHPC10]
Number of hashtags	Number of hashtags in a tweet. [APG12, CC12, PZP11, POL11]
Number of mentions	Number of mentions in a tweet. [APG12, CC12, LJ13, PZP11, POL11, SHPC10]
Number of followers	Number of followers of the author. [APG12, HDD11, LJ13, LOTW13, POL11, SHPC10, WLZL12]
Number of followees	Number of followees of the author. [APG12, HDD11, LOTW13, POL11, SHPC10, WLZL12]
Presence in lists	Number of times that an author has been added to lists. [LOTW13, POL11]
Verified	If the author has a verified account. [LOTW13, POL11]
Ratio of followers over followees	Ratio $followers/followees$ or its inverse. [APG12, PZP11]
N-grams	Presence of possible n-grams in the text. Usually used together with dimensionality reduction methods. [APG12, POL11]
Number of Stop Words	Number of stop words in the tweet. [APG12]
Time	Time when the user received the tweet. [APG12, LJ13]
Day of week	Day of the week when the user received the tweet. [APG12]
Time zone	If the author and the receiver of a tweet are in the same time zone. [LOTW13]
Wait time	Average time a user takes to reply or retweet a message. [CC12, HDD11]
Timeline position	How many messages on average a user receives between receiving and replying (or retweeting) a tweet. [CC12]
Tweet age	When the tweet being retweeted was originally created. [CC12, HDD11]
Previous interaction	If the user has already replied to or retweeted the author in the past. [CC12, LOTW13, WLZL12]
Author's activity	Absolute number, frequency, or distribution that represents how the author tweets. [CC12, HDD11, LJ13, LOTW13, PZP11, POL11, SHPC10, WLZL12]
Followees activity	Absolute number, frequency, or distribution that represents how the followees of the user tweet. [PZP11]
Tweet size	Number of characters of the tweet. [CC12, POL11]
Author's PageRank	PageRank of the author. [HDD11, WLZL12]
Reciprocal links	If the author and the user follow each other. [HDD11, PZP11, WLZL12]
Reciprocal followers	Number of followers that the author and the user share. [PZP11, WLZL12]
Reciprocal followees	Number of followees that the author and the user share. [PZP11, WLZL12]
Reciprocal mentions	Number of tweets where the author mentions the user or the user mentions the author. [PZP11]
Reciprocal retweets	Number of retweets that the author and the user share. [PZP11]
Clustering coefficients	Clustering coefficients of the network structure. [HDD11]
Previously retweeted message	If and how many times a message has been retweeted by other users in the past. [HDD11, SHPC10]
Author's retweet count	How many messages of the author have been previously retweeted. [HDD11, PZP11]
Emoticons	If there is an emoticon in the tweet. [LJ13]
Message topic	Topic identification on the message text or topic similarity measures between the author's interests and the message topic. [LJ13, LOTW13, PZP11, WLZL12]
Language	User's profile language. [POL11, WLZL12]
Favorite	If the tweet has been marked as a favorite by the author. [POL11, SHPC10]
Response	If the message received is an answer to a previous message. [POL11]
Account age	Age of the tweet author's account. [SHPC10, WLZL12]
Trending topics words	If the tweet has *trending topics*' terms. [POL11]
Reciprocal hashtags	Number of hashtags in common that the author and the user shared in the past. [WLZL12]
Reciprocal URLs	Number of URLs in common that the author and the user shared in the past. [WLZL12]
Number of lists	Number of lists that an author created. [WLZL12]

turn, affect their observed behavior. This interaction with the state of a community can confound conclusions about people's behavior, because the differences one observes may simply be due to changes in the community, rather than any significant change in the outcome variable of interest or the user population.

2.2.2 Cohorts are analytically useful

To prevent such confounding, a common unit of analysis to control for such biases is cohort analysis, widely used in fields such as sociology [MF12, Gle05], economics [AW93, Bel05], and medicine [HECR96, DAE+10]. A cohort is defined as a group of people who share a common characteristic, generally with respect to time. For example, people born in the same year, or those who joined a school at the same time, or got exposed to an intervention at similar times can be considered as cohorts. Such people in a cohort can be assumed to be exposed to the same state of the world and thus are more comparable to each other than people in other cohorts.

For example, sociological studies often use students who join a school in the same year to understand the effect of interventions [Goy08, AHP12], and condition on the year in which people were born to understand people's behavior, such as variations in financial decisions-making [AW93] or opinions on issues [FD88, Jen96]. Similarly, medical studies interpret effects of drugs using cohorts of people with the same age group or lifelong exposure to correlated conditions [HECR96, DAE+10].

Recent work shows that the importance of cohorts transfers to online communities as well. Just as people's behavior varies according to their biological age, their experience in an online community may vary with their age in the community and their year of joining. In Wikipedia, for example, we find substantial differences in the activities of cohorts of users who joined earlier versus those who joined later [WCK+11]. Similarly, on review websites, users who join later tend to adopt different phrases than the older users who had joined earlier [DNMWJP13].

2.2.3 What might cause these differences?

These differences in activity between cohorts may be due to a number of reasons. It could be due to selection effects: people who are enthusiastic about a community or its goals are more likely to self-select as early members of a community, while others may be more likely to join later [LH08].

The norms in community may change over time, which could explain why users in later cohorts may behave differently. In many cases, it is a bottom-up process. Kooti et al. [KGM10] showed that social conventions can define the evolution of a community and the early adopters play a major role in designing these conventions, even if at the time this is not known by them. Examples include adoption of 'RT', a retweeting norm by Twitter users and the subsequent introduction of the Retweet button on Twitter [KGM10]; change in language use by new and old users on review websites [DNMWJP13]; and assumptions of clear roles and responsibilities on Wikipedia [KSPC07]. In other cases, it may be directed by the community managers. For instance, the makers of Digg unilaterally changed the nature of the community by introducing a new version of the website, leading to a sudden change in norms and behavior in the community [Ing14, Lar14].

The growth of a community may also affect people's behavior. Successful communities often grow very rapidly, which can be both good and bad for people's experience with the community. On one hand, growth would imply availability of a larger chunk of content to choose from. On the other, it might be harder to connect to others and get responses in a bigger community. A community may also need to adopt new rules and policies to manage growth and newcomers, as in the evolution of Wikipedia [CAKL10, BBF+05], and in those cases, the experience of later cohorts of users may be vastly different from the initial ones who joined before formal rules were in place.

Finally, patterns of use may change because the overall population of Internet users is still changing. As more and different people become connected with the web, their influx may lead to observed change in activity patterns. This also affects technology use: people who did not grow up

in a technological environment differ in their social media and search usage compared to younger generations[CHdZ10, Bel05].

All of the above reasons suggest that users from different cohorts are likely to be different, which has also been demonstrated in online and offline communities [Ryd65, DNMWJP13, Pre01, CHdZ10]. Accounting for these differences can be helpful for making conclusions about outcomes of interest, such as user's activity levels, their survival, among many other possibilities.

Chapter 3

Using Text Similarity to Detect Social Interactions not Captured by Formal Reply Mechanisms

3.1 Reaction Identification

This section presents the definition of the problem and the method used to attack the identification of non-explicit reactions in Twitter.

When users decide to post a message in Twitter, they might be reacting to some content they saw from one of their followees. The first assumption is that the evidence for these reactions are the textual features in a given tweet by user u and textual features in the set of recent tweets by u's followees. Another assumption is that, if u tweeted in reaction to a followee's message, there should be higher text similarity between that tweet and that message. This work focus on text features, rather than user or network characteristics found in prior work, because they have been shown to be useful while simplifying data collection, computation, and modeling.

This leads to this work's first research question, about whether text similarity has potential for identifying non-explicit responses. Do explicit responses in fact tend to have high text similarity? If so, what fraction of high-scoring tweets are non-explicit? And, even when similarity is lower, when might non-explicit responses be present?

The second research question asked is how these non-explicit responses are distributed among users. Are many users "invisible" because, although they appear to be responsive based on scores, their responses are not explicit? Why are they lost? Are they naive or low-frequency users who do not know better than to retype or cut and paste or restate? And, is this likely to be important in estimating the overall responsiveness of users?

3.1.1 Influence Window

The information Twitter presents to a user is the set of tweets sent by their followees in reverse chronological order. Comarela et al. [CC12] study how far back in the user feed is a tweet when replied or retweeted. They divided the users into four sets of increasing levels of activity and found that over 80% of replies and 60% of retweets are responses to one of the 50 most recent tweets in a user's feed. They also present cumulative distributions of these replied and retweeted tweets when varying the position in the feed, and the last 100 tweets in the feed contain more than 80% of the tweets in these distributions. Based on this, a window w_i for the tweet t_i is defined as the last $n = 100$ tweets generated by user's u followees f_i immediately before t_i, taken in reverse chronological order. Figure 3.1 illustrates the window.

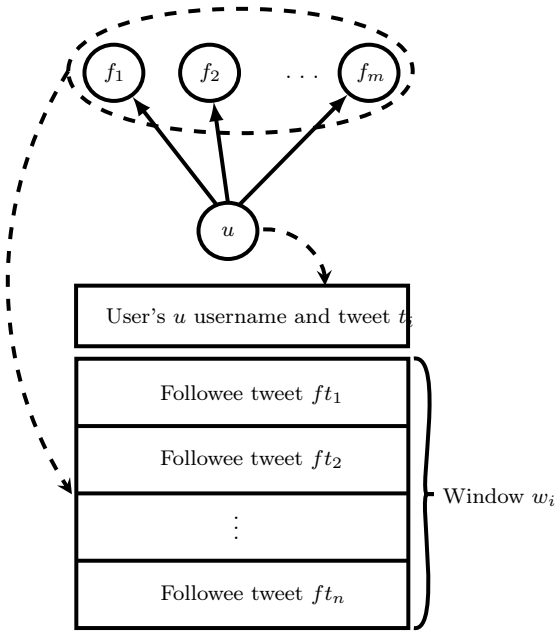


Figure 3.1: *Construction of the window w_i for the tweet t_i. The tweets in the window (in this paper, $n = 100$) are those most generated by user u's followees most recently before t_i.*

Table 3.1: *Regular expressions used to extract features from tweets.*

Hashtags	`(?:[\s|^])(#[\w]+)`
Users	`\B(?:[@?])([\w]{1,20})`
Words	`(?:^|[\s][^@?#\s\w]*)([\w]+)`

3.1.2 Textual Features

Each tweet in a user's feed also carries associated meta-data besides the message itself, such as the author's profile and user name, tweet creation time, number of times liked, and number of times retweeted. In this analysis, users are modeled as primarily paying attention to the textual content when considering a response; thus, only textual features the user's feed exposes are considered.

Tweets are first preprocessed using Python's NLTK package [BKL09] to be lower case, remove stopwords, and apply Snowball stemming, all common practices when using tf-idf scoring. Hashtags, usernames, and processed words are then extracted using the regular expressions shown in Table 3.1. Finally, the tweet author's username is added as a feature since that is also visible in the feed.

3.1.3 Message Scoring

The text similarity metric used for this task was the tf-idf scoring. It is a proven technique for information retrieval commonly employed in analyzing Twitter data. tf-idf stands for term frequency and inverse document frequency. This method takes as input a set of documents D, where each *document* is a set of *terms*, and produces a document-by-term matrix of tf-idf scores. These functions can be scaled, but usually the tf is not scaled and the idf is logarithmically scaled. For a given $(document, term)$ matrix entry, the tf function is the *term* occurrence count in the *document* and the idf function is given by Equation 3.1.

$$idf(D, term) = \log \frac{|D|}{|\{d \in D | term \in d\}|} \tag{3.1}$$

Notice here that the idf is a function of the whole set of documents and a particular *term*, while the tf is a function of the document and the *term*. One high level interpretation of these

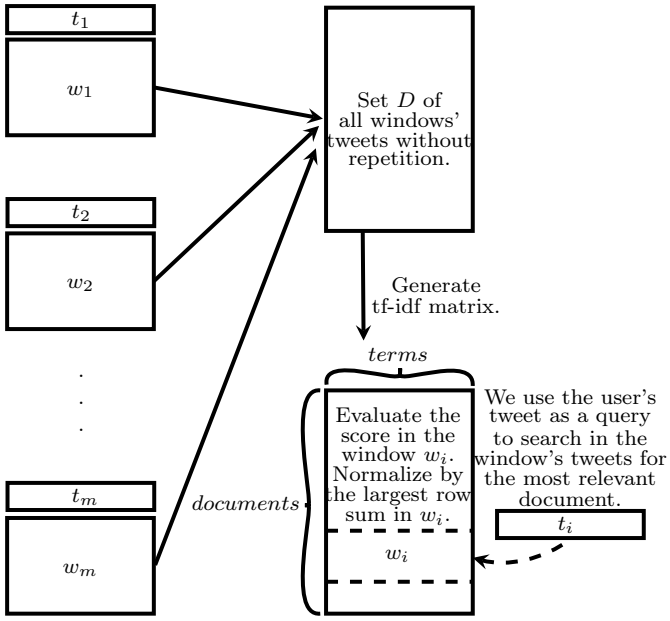


Figure 3.2: *Process to generate each tweet score. All the tweets in the windows are used to compose the corpus from which the tf-idf matrix for a given user is generated. Each user's tweets are then used as queries to search in their windows for the most relevant followee's tweet.*

functions is that tf indicates how important is *term* for the *document*, while the *idf* captures how common is the *term* among the *documents* and indicates how much information it provides when it occurs in a particular *document*.

The *tf-idf* was calculated using the implementation provided by the Python package scikit-learn [PVG+11]. It uses a smoothed version of the *idf* function (even if the *term* happens in all documents it will not be ignored). The final *tf-idf* document-by-term matrix is given by Equation 3.2.

$$tf\text{-}idf(document, term) = tf * (1 - idf) \tag{3.2}$$

The set of documents D is comprised of the tweets in all windows for a user u (each user has its own set D, and words in these tweets form a user-specific language model). Each textual feature is one *term* in our analysis, and the *tf-idf* scores matrix is computed for D.

The tweets generated by u are then used as queries that leverage the matrix. For each tweet t_i, its text features are extracted (removing duplicate *terms*) and the *score* evaluated for each pair (t_i, ft_j), where ft_j is a followee's tweet in t_i's window w_i. The *pairScore* is given by Equation 3.3.

$$pairScore(t_i, ft_j) = \sum_{term \in (t_i \cap ft_j)} tf\text{-}idf(ft_j, term) \tag{3.3}$$

To be able to compare in a score-independent way between tweets and users, the score for each tweet is normalized based on the maximum value of the *tf-idf* matrix row sum for the tweets in window w_i, as given by Equation 3.4.

$$normalization(w_i) = \max_{t \in w_i}(pairScore(t, t)) \tag{3.4}$$

This normalization means that the tweet t_i generated by the user will have a final score of 1 if that tweet reproduces the exact text of the tweet that would yield the maximum score that is present in the window w_i. The *score* for each tweet t_i is then given by Equation 3.5.

$$score(t_i) = \max_{ft_j \in w_i} \frac{pairScore(t_i, ft_j)}{normalization(w_i)} \tag{3.5}$$

The interpretation of the $score(t_i)$ is how likely t_i is to be a response to a friend's tweet $ft = argmax_{ft' \in w_i}(pairScore(t_i, ft'))$.

3.2 Twitter Dataset

This definition of potential response allows ego networks to be collected rather than full network data. This is often a more feasible approach when dealing with online social networks, since even friendly APIs normally impose rate limits. Ego networks are often useful for studying interaction and influence [WCK+11, SGC13]; here, they are appropriate because the method requires only a user's content and his followees' in order to reconstruct the feed windows.

The dataset this paper is based on was collected as part of a project to investigate differences in online behavior between political groups, driven by observations that, in the U.S. 2012 presidential election, Democrats were more active and effective in social media than Republicans. This paper draws on that dataset, using ego networks on Twitter belonging to users that followed Barack Obama crawled in the first three weeks of December 2012 using V1.0 of the Twitter API.

The crawler first got all the followers for Obama's account, then filtered out users that did not choose English as their profile language or had no tweets in the last month. It then randomly selected 547 users and collected up to one month (or the Twitter limit of 3200 historical tweets) of Tweets from each user and all of their followees, creating a set of ego networks.

Because of the 3200 tweet per-user limit, as well as occasional API or network errors, the dataset does not contain a complete record of all followees' tweets. This could affect estimates of the presence of non-explicit responses; thus, networks where a significant proportion of followees' tweets appeared to be missing were filtered out. Tweets were considered missing when a followee's activity only partially overlapped with the ego user's[1], with the number of missing tweets estimated based on the length of overlap and the rate of that followee's tweets. Users for whom over 20% of their followers' tweets were estimated missing were removed from the dataset, leaving 449 ego networks[2].

3.3 Results

3.3.1 Overview

The methodology described to extract the windows was applied for all 449 ego networks, computing windows for each tweet an ego user u authored within 14 days of their most recent tweet in the dataset. Table 3.2 provides an overview of the dataset and the generated windows, while Figure 3.3 presents the cumulative distribution of the time length for the generated windows. Tagged tweets are defined as those indicated by the API as explicit responses, i.e., Replies and Retweets, while the Non-Tagged set is anything not tagged by Twitter[3].

3.3.2 How prevalent are non-explicit responses?

This section addresses the first research question of whether or not text similarity has potential for identifying untagged responses, starting with whether Tagged reactions indeed tend to have higher scores than Non-Tagged ones.

Mean and median scores are lowest for Non-Tagged and highest for Retweets, as shown in Table 3.3. This can also be seen in the scores' histogram for each of these sets in Figure 3.4. The

[1] There is a parallel, opposite problem for users who added followees during the ego user's activity period; windows for tweets before the followee was added will incorrectly contain their tweets, which the user could not have responded to. We saw no good way to address this and so tolerate the error.

[2] Other thresholds (5%, 10%, 50%, 80%, 100%) were tested. Lower values lead to similar results, while higher values increased the number of users that lacked data for analysis; 20% was chosen as a reasonable trade-off between sample size and meaningfulness of results.

[3] Upper-case names refer to the collected sets in this work, while lower-case names refer to messages in general.

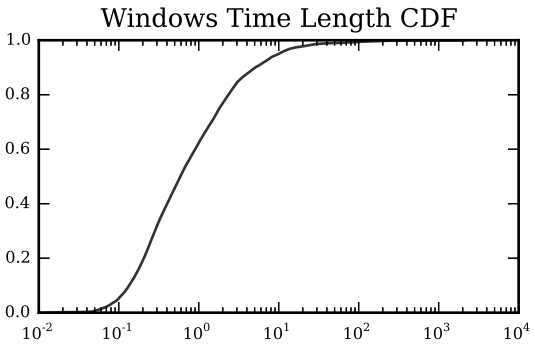


Figure 3.3: *Cumulative distribution function for the time length of the windows given in hours. Most windows' lengths are in the interval* $[10^{-1}, 10^{1}]$ *hours; about 60% of windows are 1 hour or less, meaning users receive on average over 100 tweets an hour.*

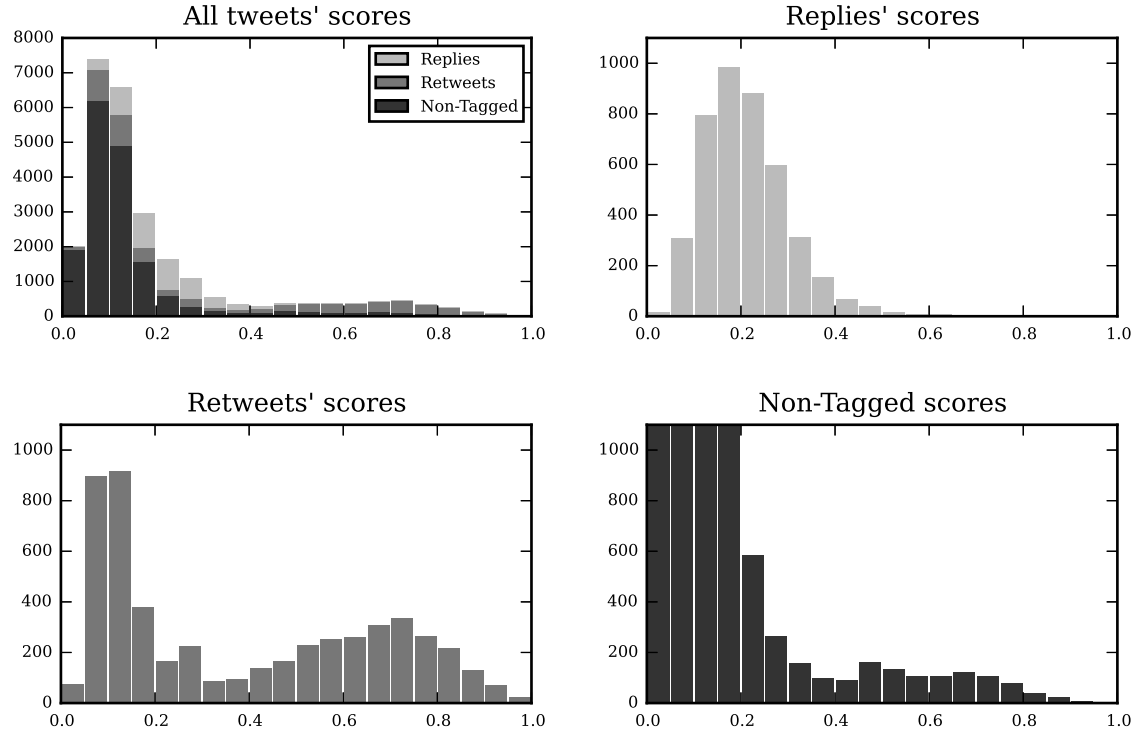


Figure 3.4: *Histograms for the normalized similarity scores. Note that the y-axis for the Non-Tagged subgraph was truncated at 1100 for better visualization of the tail of the distribution and matching other scales. Retweets have a higher average score than Replies, which in turn are higher than Non-Tagged. Further, Retweets have a bimodal distribution; high scores are near-duplicates of the tweets they are responding to, but over 54% have a score below 0.384, suggesting that people often substantially edit retweets or retweet items not in their feed windows.*

Table 3.2: *Descriptive data. Tweet counts are based on ego users. Each tweet may be tagged either as a retweet or a reply. Replies also provide the replied tweet id, allowing us to count how often a tagged reply refers to a tweet in the window. As with Comarela et al. [CC12], over 80% of tagged replies reference one of the 100 most recent tweets.*

Users	449
Tweets	26051
Average Tweets/User	58.02
Min Tweets/User	1
Max Tweets/User	832
Retweets	5209
Replies	4192
Replies in windows	3455
Window avg. size (h)	5.24
Windows std. deviation (h)	63.87
Windows min size (h)	0.01

Table 3.3: *Sample mean and standard deviation for the normalized similarity score for the Replies, Retweets, and Non-Tagged sets.*

	Mean	Median	Std.
Non-Tagged	0.135	0.102	0.136
Replies	0.212	0.200	0.092
Retweets	0.384	0.287	0.282

Table 3.4: *Number of high scored messages and the total of messages for the sets Non-Tagged, Replies and Retweets. The highlighted number of high scored Non-Tagged messages is around 11% of the highlighted total of Tagged messages.*

	Non-Tagged	Replies	Retweets
High Scored ($score \geq 0.384$)	998	177	2408
Total	16650	4192	5209

score behaves as expected when we consider the averages, returning higher values for Replies and Retweets. However, the proximity of the means for the Replies and Non-Tagged and the higher variance of the Non-Tagged makes these two distributions not so well distinguishable based on score alone. The Retweets, on the other hand, present a heavier tail on the distribution. This suggests that the score captures general trends of the Tagged tweets, but is more suitable for Retweets. Considering that the Retweet average is 0.384 and that it is higher than the Replies mean by more than one standard deviation, **high scored messages** are defined as messages with $score \geq 0.384$.

Although the Non-Tagged set has a lower average, it has a higher variance than replies. This comes from the fact that Non-Tagged tweets have a heavier tail when compared to replies, as seen in Figure 3.4. Also, the Non-Tagged high scored tweets are not neglectable when compared with the number of high scored Tagged tweets, as seen in Table 3.4: such Non-Tagged tweets would comprise about 11% of responses, even with a fairly conservative cutoff of 0.384. However, high scored messages misses most of the explicit Replies with this cutoff choice.

Considering the retweet behavior, it would be expected that the normalized similarity score for retweeted messages would be high as long as the original tweet showed up in the windows and the retweet is basically reproducing the message with almost no modifications. Surprisingly, this is not what is observed in Figure 3.4. Instead, more than 54% of Retweets have a $score < 0.384$. One possible explanation for this is that people sometimes retweet when they use other parts of the interface, such as other users' profiles or search results, or use social media share buttons attached to tweets on other sites. Another possibility is that people might frequently edit retweets.

3.3.3 Features of Replies, Retweets and Non-Tagged messages

To help understand the mystery of low-scoring retweets, and more generally to understand what sorts of markers the method is using to identify potential responses, a sample of representative tweets from each category across a range of normalized similarity scores is examined. Table 1 (see the Appendix) shows both the user's tweet (top in each pair) and the text of the highest-scoring followee's tweet in the window for that tweet (bottom in each pair).

For system tagged Retweets, most of the high scored content has almost the same content as the original message (as expected), as in tweets #1 and #2 in the table. One interesting thing to notice here is that as the tweet length decreases, the normalized similarity score goes down (compare #6 to #1). This is related to the fact that the *tf-idf* score is sensitive to the number of matched words between the query and the document. Below a threshold of around 0.3 in this dataset, this effect disappears. Instead, the text starts to look more like two tweets about a common external topic (#7, #8, #9)—despite the fact that the tweet text preserves the "RT" retweet marker. These would be likely candidates for actual retweets that occur outside the window, either farther back in the feed or other parts of the interface than the feed.

When looking at system tagged Replies, high-scoring replies show two main patterns. In one, they look largely like retweets that were tagged as replies, likely because people pressed the reply button and pasted text from the text they replied to, as in #11. In the other, the tweet mentions multiple users who are conducting an ongoing conversation and want all of them to be notified when someone posts something new, as in #12 and #13. It is important to notice that this set of tweets has a maximum score lower than the other sets; scores on the higher end of the distribution could not be found. Also, it appears that @-mentions are the main source of evidence for the normalized similarity scoring even as it goes down, and in fact, replies with low scores still often look like replies despite the low *tf-idf*. This is often (#16, #19) but not always (#18, #20) indicated by bi-directional @-mentions of the conversational partner.

When looking at Non-Tagged tweets, one of the first things to notice is that high scored tweets usually are retweets that were not captured by the system. In some cases it is likely users are manually copying the content of the messages and adding retweet markers (#23, #24); in others, it is more likely that both users are independently retweeting external content (#21, #22). Users often make small comments together with the original text (#22, #23, #25). As the normalized similarity score goes down, the messages look less like a retweet, but often still appear to be topically related, sometimes via hashtags (#28, #29).

In general, higher normalized similarity scores seem to capture retweets reasonably well, even though being sensitive to their length, and a particular type of reply that involves conversations. Non-tagged tweets with high scores are often retweets or quotes with extra comments from the users, although sometimes the retweets may be common retweeting of external content rather than retweets from the window. Further, even the conservative estimate chosen shows that non-explicit responses are quite common—and it is likely that a number of the of the "middle scoring" tweets are actual responses. Distinguishing those from external influences or underlying interest similarity would be an important next problem in building better models of non-explicit response.

3.3.4 Variations in User Responsiveness

The previous sections demonstrate that it's likely that 11% or more of Non-Tagged tweets are responses that are not explicitly captured by the system. This section addresses the other main research question of how these losses are distributed among different users in the network.

These Non-Tagged high-scored messages were authored by 129 of the 449 users (29%). This suggests that users generate responses that are missed in a non-uniform way: many users behave as the system expects, using explicit reply and retweet mechanisms, but a significant number respond, at least sometimes, without using those mechanisms. Figure 3.5 shows histograms for example users that have most or all of their responses untagged by Twitter even though they present a high *score*. Note that these users span a range of activity levels, meaning that they are not just newbies that

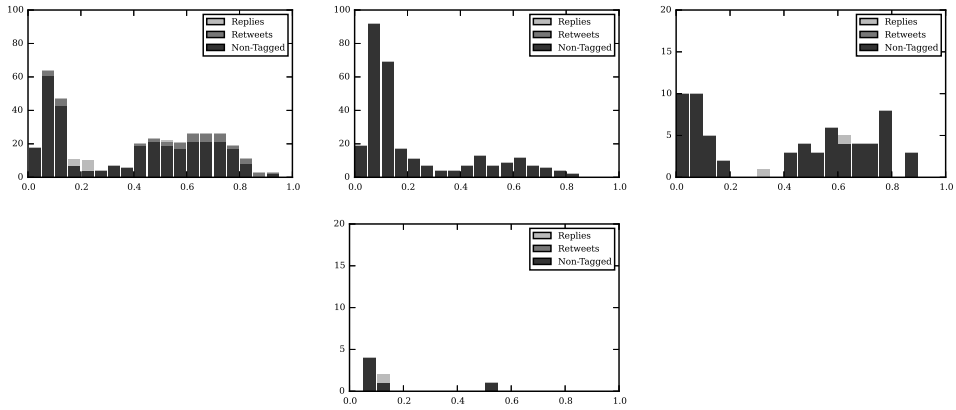


Figure 3.5: *Score histograms for sample users who present a significant amount of high scored Non-Tagged content relative to their total amount of messages, which indicates that most of their reactions are not being properly tagged by Twitter.*

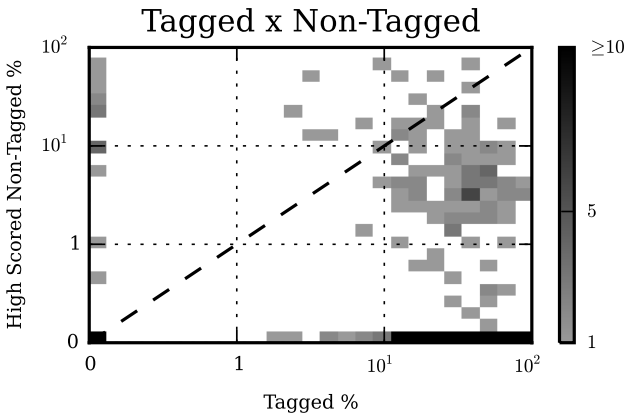


Figure 3.6: *2D histogram of the percentage of Tagged and high-scored Non-Tagged messages for all users. The scale is linear in the interval* $[0, 1]$ *and logarithmic on the interval* $(1, 100]$*; the dashed line represents an equal percentage of Tagged and Non-Tagged tweets. Many users are non-responsive (the point at the origin) or use the explicit response mechanisms consistently (points hugging the x-axis with a 0 value for high scored Non-Tagged %). However, a significant number never use the explicit response mechanisms (points hugging the y-axis with a 0 value for Tagged %), use them only occasionally (points above the dashed line), or occasionally forget to use them (points below the dashed line).*

don't know how to use the interface.

In order to better understand the behavior distribution among all users, Figure 3.6 shows a 2d-histogram for the points (p_i^T, p_i^N), where each of these points is the percentage of the Tagged messages p_i^T and the percentage of the high scored Non-Tagged messages p_i^N. Each of these points is evaluated for a user u_i in relation to the total number of messages the user authored. The high scored Non-Tagged percentage p_i^N is the proportion of this user behavior that were likely to be reactions while the percentage p_i^T is the proportion of reactions actually captured.

The 111 users that never have messages that scored higher than 0.384 nor used explicit system reply mechanisms are concentrated at the origin of the histogram. Users that lay on the x-axis only react through explicit reaction mechanisms the system offers, therefore have all their reaction Tagged. Similarly, users on the y-axis never use explicit reaction mechanisms, although they present high scored Non-Tagged content. Users above the dashed line have more high scored Non-Tagged content than Tagged content. It is possible to say that users that lay above the dashed line are more likely to produce content that can be missed by Twitter's tagging system, and they account for 27 users, about 6% of the dataset.

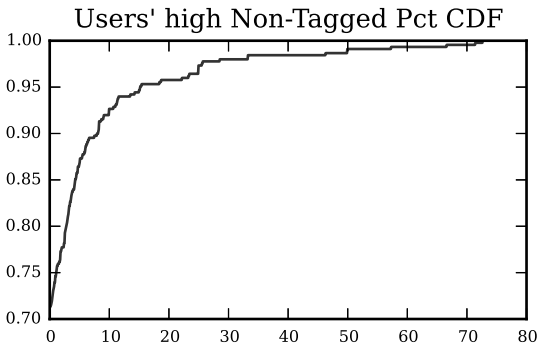


Figure 3.7: *Cumulative distribution of the users for the percentage of high scored Non-Tagged messages. 71% of the users have no high scored Non-Tagged messages, while 8% of the users had at least 10% of their messages high scored and Non-Tagged.*

When considering the cumulative distribution of the users according to the percentage of high scored Non-Tagged messages p_i^N, shown in Figure 3.7, we identify more than 8% of the users with at least 10% of their messages being high scored and missed by Twitter's tagging system.

These results indicate that methods that rely on explicit indicators of response likely miss or seriously under-represent the behavior of a sizable proportion of the Twitter population.

Chapter 4

Evolution of effort and activity in online communities from a cohort perspective

4.1 Data: Reddit as a community

In this paper, we explore how cohort based analysis can give us deeper insights into three common questions about online communities: how active are users [SM11, HP09, JSFT07, LW84], how much do they contribute [SM11, GLNGT04, GTC+09], and what kinds of work do they engage in [WCK+11, CAKL10, PHT09]? We do this in the context of Reddit, beginning with a brief overview of both Reddit and the dataset that we use in this paper, focusing on aspects that directly impact our analyses[1].

4.1.1 What is Reddit, briefly

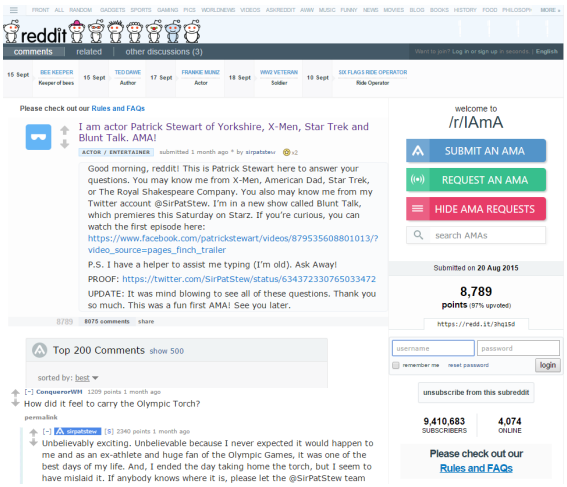


Figure 4.1: *Reddit interface for visualizing a submission. Here we see a Patrick Stewart "IAmA", when he is online answering questions in the comments of his submission. We can see the most upvoted comment and Patrick's answer right below.*

Reddit is one of the largest sharing and discussion communities on the Web. According to Alexa, as of late 2015 Reddit is in the top 15 sites in the U.S. and the top 35 in the world in terms of monthly unique visitors. It consists of a large number of subreddits (853,000 as of

[1]There is more to say about Reddit itself (see https://www.Reddit.com/about/).

June 21st, 2015[2]), each of which focuses on a particular purpose. Many subreddits are primarily about sharing web content from other sites: in "Pics", "News", "Funny", "Gaming', and many other communities, users ("Redditors") make "submissions" of links posted at other sites that they think are interesting. In other subreddits, Redditors primarily write text-based "self-posts": "AskReddit", "IAmA", "ShowerThoughts" are places where people can ask questions and share stories of their own lives. Generically, we will refer to submissions and text posts as "submissions".

Each post can be imagined as the root of a threaded comment tree; in addition to submitting, Redditors can make comments, and vote on both submissions and comments. Votes are used both to sort comments within a submission and submissions within a subreddit, and also form the basis of "karma", a reputation system that essentially tracks how often people upvote a given Redditor's comments and submitted links. We can observe these elements in Figure 4.1. Redditors can also create subreddits and volunteer to moderate them.

We choose Reddit as our target community for a number of reasons. It has existed since 2005, meaning that there has been ample time for the community to evolve and for differences in user cohorts to appear. Second, it is one of the most popular online communities, allowing different types of contributions—comments and original submissions—across many different subreddits.Third, Reddit data are publicly available through an API.

4.1.2 The dataset

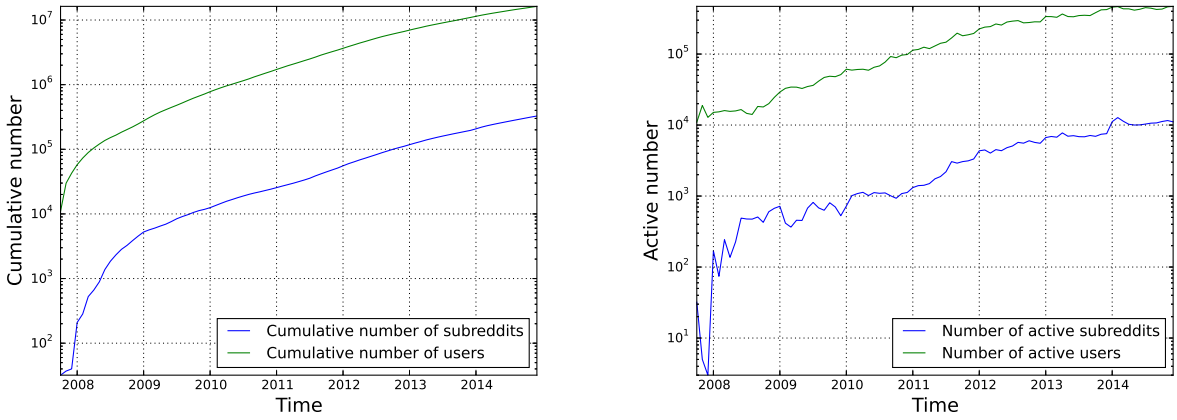


Figure 4.2: *Figure (a) shows the cumulative growth of Reddit for users and subreddits. Figure (b) shows the number of active users and subreddits in Reddit over time. An active user or subreddit is one that had at least one post (comment or submission) in the time bin we used—here, discretized by month.*

Redditor *Stuck_ In_ The_ Matrix* used Reddit's API to compile a dataset of almost every publicly available comment[3] from October 2007 until May 2015. The dataset is composed of 1.65 billion comments, although due to API call failures, about 350,000 comments are unavailable. He also compiled a submissions dataset for the period of October 2007 until December 2014 that was made available for us upon request, containing a total of 114 million submissions. These datasets contain the JSON data objects returned by Reddit's API for comments and submissions[4]; for our purposes, the main items of interest were the UTC creation date, the username, the subreddit, and for comments, the comment text.

[2] http://www.Redditblog.com/2015/06/happy-10th-birthday-to-us-celebrating.htmlfor more numbers on Reddit size.

[3] Available in https://www.Reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_Reddit_comment.

[4] A full description of the JSON objects is available at https://github.com/Reddit/Reddit/wiki/JSON.

We focus on submissions and comments in the dataset because they have timestamps and can be tied to specific users and subreddits, allowing us to perform our time-based analyses. In some analysis, we look only at comments; in some, we combine comments and submissions, calling them **"posts"**. We would also like to have looked at voting behavior as a measure of user activity[5], but individual votes with timestamps and usernames are not available through the API, only the aggregate number of votes that posts receive.

4.1.3 Preprocessing the dataset

To analyze the data, we used Google BigQuery[6], a big data processing tool. Redditor *fhoffa* imported the comments into BigQuery and made them publicly available[7]. We uploaded the submission data ourselves using Google's SDK.

For the analysis in the paper, we did light preprocessing to filter out posts by deleted users, posts with no creation time, and posts by authors with bot-like names[8].

We also considered only comment data from October 2007 until December 2014 in order to have a matching period for comments and submissions. After this process, we had a total of 1.17 billion comments and 114 million submissions.

4.1.4 An overview of the dataset

Here we present an overview of the dataset that shows Reddit's overall growth. Figure 4.2a presents the cumulative number of user accounts and subreddits created as of the last day of every month. After an initial extremely rapid expansion from 2008–2009, both the number of users and subreddits have grown exponentially. As of the end of 2014, about 16.2 million distinct users have made at least one post and 327,000 subreddits received at least one post based on our data.

However, as with many other online sites, most users [SM11, HP09, JSFT07] and communities [ABJ06] do not stay active. We define as an "**active user**" one that made at least one post in the month in question. Similarly, an "**active subreddit**" is one that received at least one post in the month. In December 2014, about 470,000 thousand users and 11,400 subreddits were active, both an order of magnitude less than the cumulative numbers. Figure 4.2b shows the monthly number of active users and subreddits.

Our interest in this paper is not so much whether users survive as it is about the behavior of active users. Thus, in general our analysis will look only at active users and subreddits in each month; those that are temporarily or permanently gone from Reddit are not included.

4.1.5 Identifying cohorts

We define the "**user's creation time**" as the time of the first post made by that user. Throughout this paper, we will use the notion of user cohorts, which will consist of users created in the same calendar year.

In many cases, we will look at the evolution of these cohorts. Since users can be created at any time during their cohort year, and our dataset ends in 2014, we are likely to have a variation on the data available for each user of up to one year, even though they are in the same cohort. To deal with this, some of our cohorted analyses will consider only the overlapping time window for which we collect data for all users in a cohort. This means that we are normally not going to include the 2014 cohort in our analyses.

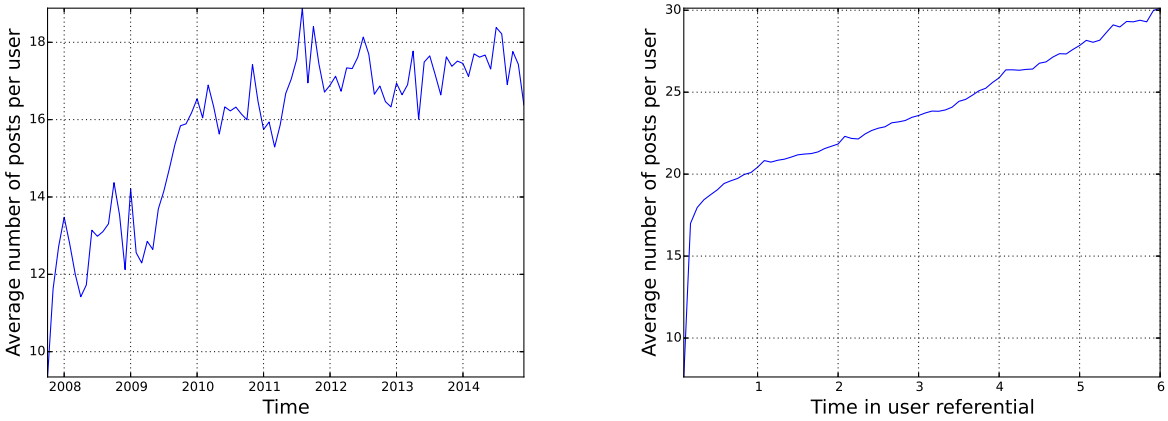
Our data starts in October 2007, but Reddit existed before that. That means that, not only do we have incomplete data for the 2007 year (which compromises this cohort), but there might also be users and subreddits that show up in 2007 that were actually created in the previous years.

[5] This would also give us more insight than usual into lurkers' behavior; we'll return to this in the discussion.

[6] https://cloud.google.com/bigquery/.

[7] See https://www.Reddit.com/r/bigquery/comments/3cej2b/17_billion_Reddit_comments_loaded_on_bigquery/.

[8] Ending with "_bot" or "Bot"; or containing "transcriber" or "automoderator".



(a) **(b)**

Figure 4.3: *In Figure (a), monthly average posts per active user over clock time. In Figure (b), the monthly average posts per active users in the user-time referential, i.e., message creation time is measured relative to the user's first post. Each tick in the x-axis is one year. In both figures (and all later figures), we consider only active users during each month; users that are either temporarily or permanently away from Reddit are not included.*

Since we can not control for these, we will also omit 2007 cohort. We will, however, include 2007 in the overall analyses over time (the non cohorted ones) for two reasons: first, it does not have any direct impact in the results, only extends the axis for 3 extra months, and second, we often compare the cohorted approach with a naive approach based on aggregation, and we would not expect a naive approach to do such filtering.

4.2 Average posts per user

In this section, we will use a common metric of user activity in online communities, the number of posts per user over time. Approaches that consider the total number of posts per user in a particular dataset [GLNGT04] and that analyzes the variation on the number of posts per user over the days [GTC+09] have been applied to online social networks.

As we will see, both visualizing behavior relative to a user's join time rather than calendar time and using cohorts provide additional insight into posting activity in Reddit compared to a straightforward aggregate analysis based on clock time.

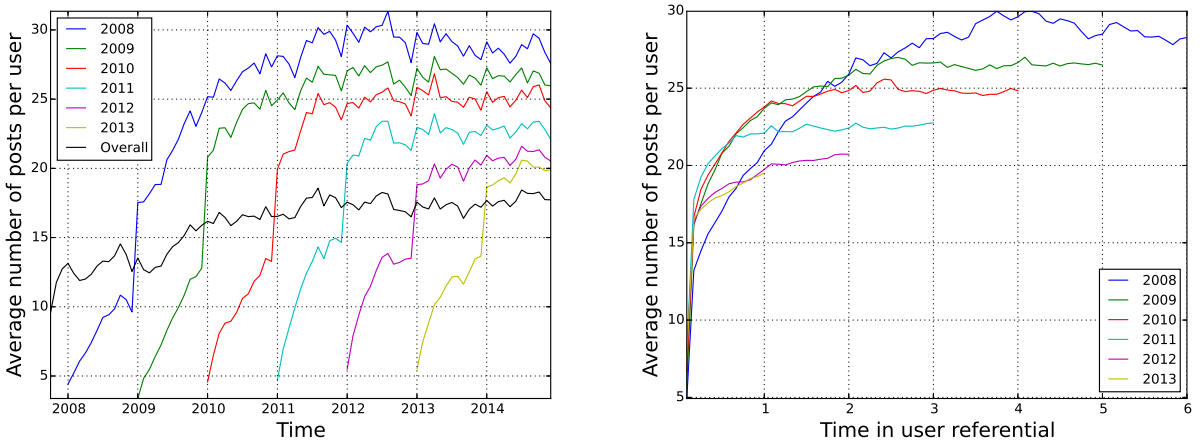
4.2.1 Calendar versus user-relative time

We start with a common analysis used in this kind of work: aggregating behavior in the community based on calendar time. Figure 4.3a shows the average number of posts per month by active users in that month. Taken at face value, this suggests that over the first few years of Reddit, users became more active in posting, with per-user activity remaining more or less steady since mid-2011.

This average view hides several important aspects of users' activity dynamics. Previous work has looked into behavior relative to the user creation time. It has been shown that edge creation time in a social network relative to the user creation follows an exponential distribution [Tom08]. User lifetime, however, does not follow a exponential distribution and some types of user content generation follow a stretched exponential distribution [GTC+09]. In Figure 4.3b, we show a different view that emphasizes the trajectory over a user's lifespan. Here, we scale the x-axis not by clock time, as in the left figure, but by time since the user's first post: "1" on the x-axis refers to one year since the user's account first post, and so on.

One caution about interpreting the graphs that are relative to the user's start time is that the amount of data available rapidly decreases over time as users leave the community, meaning that values toward the right side of an individual data series are more subject to individual variation. A tempting conclusion at this point is that the longer a user survives, the more posts they make over time. This conclusion, however, is incorrect; we will present a more nuanced description of what is happening informed by cohort-based analyses.

4.2.2 New cohorts do not catch up



(a) (b)

Figure 4.4: *Figure (a) shows the average number of posts per active users over clock time and Figure (b) the active users in the user-time referential, both segmented by users' cohorts. The user cohort is defined by the year of the user's creation time. For comparison, the black line in Figure (a) represent the overall average.*

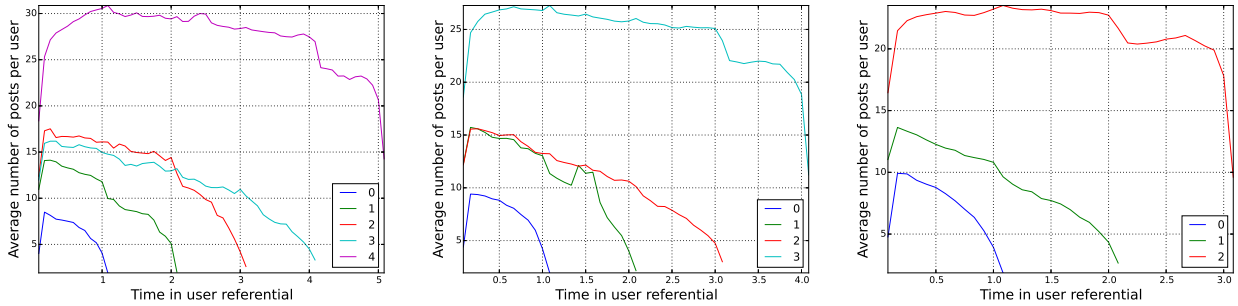
Figure 4.3b suggests that older users are more active than newer ones, raising the question of whether newer users will eventually follow in older users' footsteps. Analyzing users' behavior by cohort is a reasonable way to address this question.

Figure 4.4a shows our first attempt at this analysis. This Figure already shows a significant cohort effect: users from later cohorts appear to level off at a significantly lower posting average than users from earlier cohorts. It suggests that newer users probably will not ever be as active on average as older ones.

However, Figure 4.4a also has an awkward anomaly, the rapid rise in the average number of posts during each cohort's first calendar year, especially in December. Combing cohort segmentation with user-referential analysis, as in Figure 4.4b, helps smooth out this anomaly and aligns cohorts with each other. Doing this alignment makes clear that differences between earlier and later cohorts are apparent early on.

4.2.3 Does tenure predict activity, or vice versa?

These graphs still support the tempting conclusion that users become more active the longer they exist in Reddit, and they do not explain the very rapid increase in posting activity in the first few months. An alternative hypothesis, inspired by the "Wikipedians are Born, not Made" paper [PHT09], is that individual users come in with different posting propensities, and the rise over time is not that individual users become more active but that low-activity users leave the system. To examine this, we further segment each cohort by the number of years they were active in the system, as defined by the difference between their first and last post times.



(a) *2010 cohort* **(b)** *2011 cohort* **(c)** *2012 cohort*

Figure 4.5: *Each Figure corresponds to one cohort, from 2010 to 2012, left to right. The users for each cohort are further divided in groups based on how long they survived: users that survived up to 1 year are labeled 0, from 1 to 2 years are labeled 1, and so on. For all cohorts, longer-tenured users started at higher activity levels than shorter-tenured ones.*

Figure 4.5 shows this analysis for the 2010, 2011 and 2012 cohorts[9]. Across all cohorts and yearly survival sub-cohorts, users who leave earlier come in with a lower initial posting rate. Thus, the rise in average posts per active user is driven by the fact that users who have high posting averages throughout their lifespan are the ones who are more likely to survive. As the less active users leave the system, the average per active user increases. In other words, the correct interpretation of Figure 4.3b isn't that longer-lived users post more. It actually is that users who post more—right from the beginning—live longer.

Combining Figure 4.5's insight that the main reason why these curves increase is because the low posting users are dying sooner with the earlier observation that the stable activity level is lower for newer cohorts suggests that low-activity users from later cohorts tend to survive longer than those from earlier cohorts. That is, people joining later in the community's life are less likely to be either committed users or leave than those from earlier on: they are more likely to be "casual" users that stick around.

4.3 Comment length

Activity as measured by the average number of posts per user is one proxy for user effort. Comment length can also be considered as a proxy for user effort in the network. Users that type more put more of their time in the network, contribute with more content, and might create stronger ties with the community. Thus, we investigate how comment length has changed in the community over time, both overall and by cohort.

4.3.1 Comment length drops over time

Figure 4.6a shows the overall comment length in Reddit over time (the black line) and the overall length per cohort. Based on the downwards tendency of the overall comment length in Figure 4.6a, one might infer that users' commitment to the network is decreasing over time, or that there is some community-wide norm toward shorter commenting.

This, however, might not be the best way to interpret this information. Figure 4.6b shows the comment length per cohort based on the user referential time. An important thing to notice here is that younger users start from a lower baseline comment length than older users. Together with the fact that recent Reddit has experienced exponential growth, the weight when evaluating the overall average for Figures 4.6a and 4.6b as the years go by is shifted towards the length of

[9]We only show these figures for the sake of saving space, but the same trends are observed in the other cohorts.

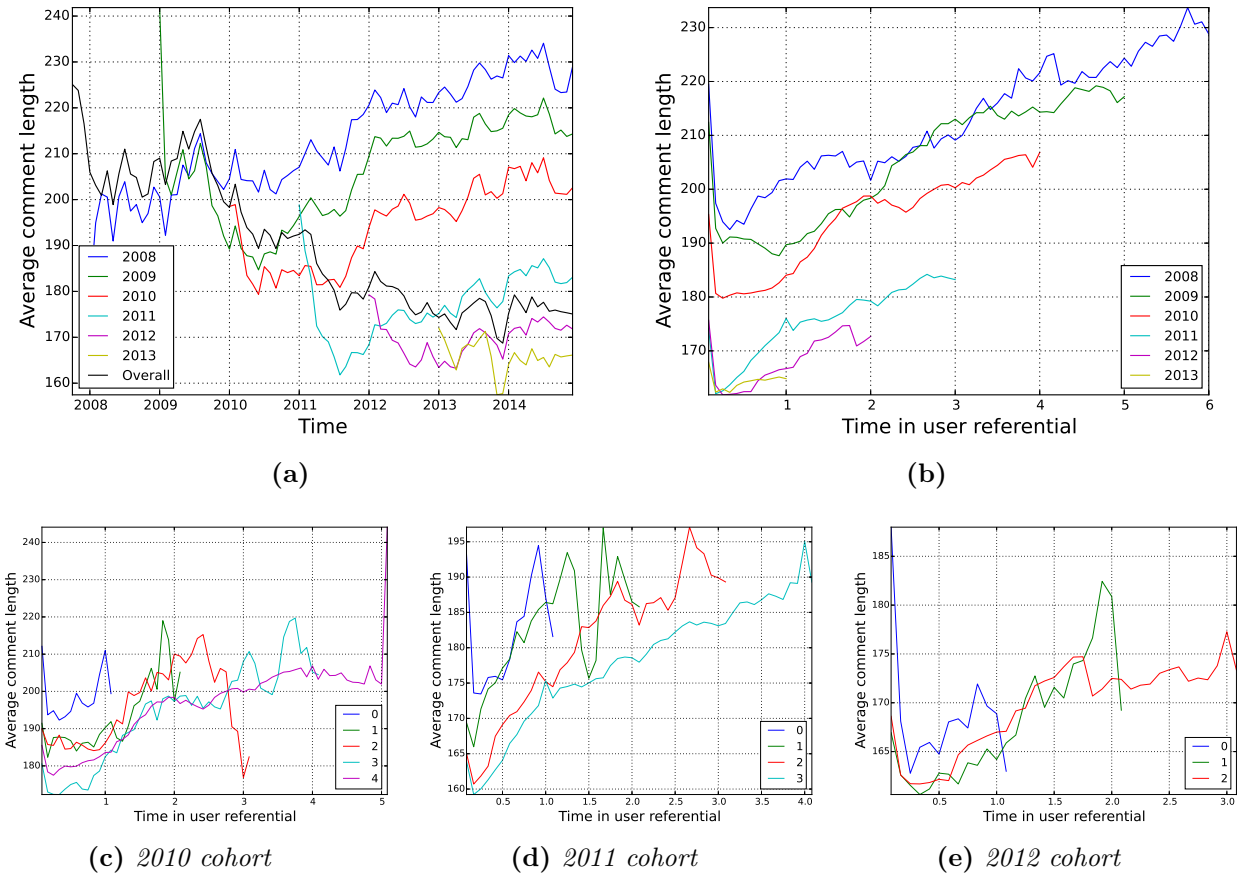


Figure 4.6: *Figure (a) shows the average comment length over clock time and Figure (b) from the user-referential time. Both figures show the cohorted trends. The overall average length per comment decreases over time, although for any individual cohort, it increases after a sharp initial drop. Figures (c), (d) and (e), similar to Figure 4.5, shows the monthly average comment length for active users in the cohorts of 2010, 2011 and 2012, segmented by the number of years that the user survived in the network. Opposite the analysis for average posts, which showed that low-activity users were the first to leave Reddit, here, people who start out as longer commenters are more likely to leave.*

the ever-growing younger generation; this younger generation brings the average down since they writing less on average.

4.3.2 Simpson's Paradox: the length also rises

Let us go back to Figure 4.6a, which shows the overall average comment length on Reddit over time. We see a clear trend towards declining length of comments in the overall line (the black line that averages across all users). This could be a warning sign for Reddit community managers, assuming longer comments are associated with more involved users and healthier discussions. A data analyst looking at these numbers might think about ways to promote longer comments on Reddit.

However, in Figure 4.6b, we saw that average comment length increases over time for every cohort. While later cohorts start at smaller comment length, after an initial drop, all cohorts show positive trends towards writing longer comments over time. This is puzzling: when each of the cohorts exhibits a steady increase in their average comment length, how can the overall mean comment length decrease? This anomaly is an instance of the Simpson's paradox [SS51], and occurs because we fail to properly condition on different cohorts when computing mean comment length.

	Cohorts								
Year	2007	2008	2009	2010	2011	2012	2013	2014	Overall
2007	220	-	-	-	-	-	-	-	220
2008	208	198	-	-	-	-	-	-	204
2009	224	204	201	-	-	-	-	-	208
2010	223	204	189	184	-	-	-	-	193
2011	233	211	199	184	167	-	-	-	182
2012	241	221	212	197	173	167	-	-	178
2013	244	225	214	199	177	167	164	-	174
2014	246	229	217	204	183	172	165	176	176

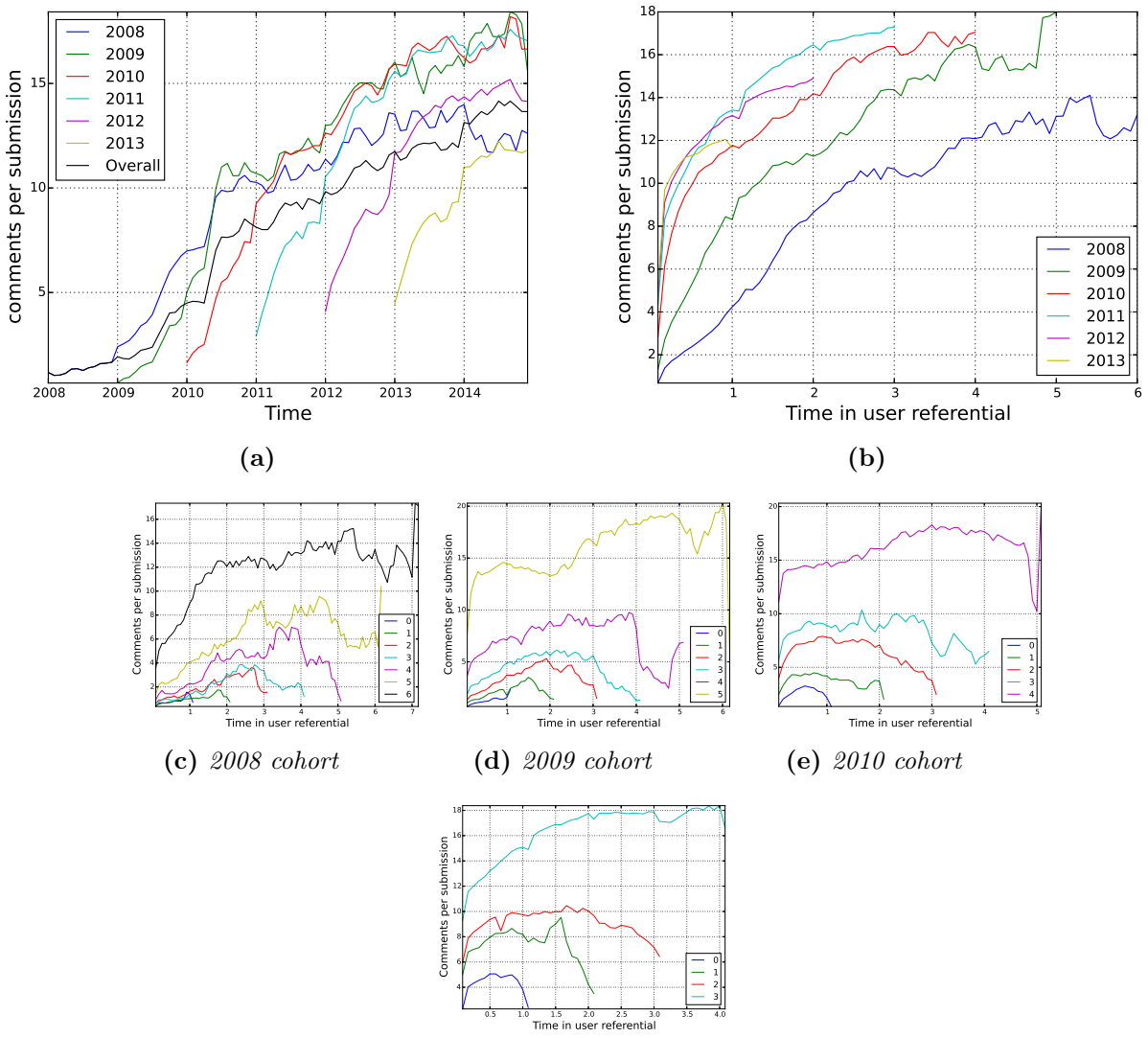
Table 4.1: *Evolution of the average throughout the years for each cohort. Each column here is one cohort and each line is one year in time. Cohorts only start having data on the cohort year, therefore the upper diagonal is blank. On the right column we see the overall average for all users.*

Table 4.1 provides some clues to what might be going on. When we move down the rows, we observe an increasing tendency in each cohort column. It means that the average comment length increases for these users. However, when we move right through the columns, people in later cohorts tend to write less per comment. If we were to average each row, we would still get an overall increasing comment length per year, but that is not what we see in the overall column. What happens here is that the latter cohorts have many more users than earlier ones. Since their numbers increase year by year, we have a much larger contribution from them towards comments, compared to users of earlier cohorts. This uneven contribution leads to the paradox we observed in Figure 4.6a.

Without the decision to condition on cohorts, one would have gathered an entirely wrong conclusion. People are not writing less as they survive, rather those who tend to write less are joining the community in much larger numbers. Knowing this, one may focus on better onboarding processes for newcomers, or try to learn why users in later cohorts tend to write smaller comments on average.

4.3.3 New users burn brighter

As with the posting per user, we can not say if the increase in the curves seen in 4.6b are due to the lower effort users dying first or because users are writing more as they live on the network. To answer this, 4.6c allow us to make two important observations: first, *comment length does increase inside of each cohort*, no matter how long the user survives. Secondly, as a general trend, *users*



(a) **(b)**

(c) *2008 cohort* **(d)** *2009 cohort* **(e)** *2010 cohort*

(f) *2011 cohort*

Figure 4.7: *Figure (a) shows the average comment per submission ratio over clock time for the cohorts and the overall average. Figure (b) shows the average comment per submission from the user-referential time for the cohorts. Figures (c), (d), (e) and (f), similarly to Figure 4.5, shows the 2008, 2009, 2010, and 2011 cohorts, segmented by the number of years a user in the cohort survived. As with average posts per month, users who stay active longer appear to start their careers with a relatively higher comments per submission ratio than users who abandon Reddit sooner. Unlike that analysis, however, the early 2008 cohort ends up below the later cohorts in Figure (b).*

that make longer comments inside of each cohort die faster. This is quite surprising, given that we would expect people to put less effort when they are more likely to stop using the network.

4.4 Kinds of contributions

One common question from the literature is what sorts of activities users engage in, for instance, to categorize users into roles they play in the community[WCK+11].

4.4.1 Over time, responsiveness increases

Consider the case of Usenet: people who never start threads and only respond play the role of answerer, while there are other roles that include fostering discussion [WGF07]. These might naturally map onto people who primarily comment and who primarily submit in Reddit, respectively. While submissions can be considered new content that an author generates, a comment can be considered as a contribution to an existing content from another author.

Since the total number of comments always surpasses the number of submissions, Figure 4.7a shows the evolution of the overall and cohorted ratio of the average number of comments a user makes for each submission they make over time from 2008 until 2013. Here we see that users who most prefer commenting to submitting come from 2009, 2010 and 2011, and we observe that, over time, the average ratio of comments to submissions increases both overall and per-cohort for active users.

Again, we analyze our data from the user-time referential, as seen in Figure 4.7b. It shows a clear pattern for users in earlier cohorts to have a lower comment per submission ratio than users in later cohorts ones, given that they both survived the same amount of time. Surviving users from later cohorts also exhibit a more rapid increase in comments per submission than those from earlier cohorts. In particular, the 2008 and 2009 cohorts increase much more slowly over time than those from 2010 onwards; later cohorts are more similar (although the 2012 and 2013 cohorts may level off lower than 2011 based on the limited data we have).

4.4.2 Comment early, comment often

Figure 4.7c shows the cohorts from 2008, 2009, 2010 and 2011 segmented by surviving year. Three interesting observations arise from these data. First, we see that just as in the analysis of average posts per user, the users who survive the longest in each cohort are also the ones who hit the ground running. They start out with a high comment-to-submission ratio relative to users in their cohort who abandon Reddit more quickly. This suggests that both the count of posts and the propensity to comment might be a strong predictor of user survival.

Second, and unlike the case for average post length, surviving users' behavior changes over time. Figure 4.5 shows that even for the most active users, they come in at a certain activity level and stay there, perhaps even slowly declining over time. Here, the ratio of comments to submissions increases over time; combined with the observation that overall activity stays steady, this suggests that the ratio is changing because people *substitute* making their own submissions for commenting on others' posts.

Finally, this increase is most pronounced for the earliest cohorts from 2008 and 2009, with ratios more than doubling over their first year, much more than the change for later cohorts. Still, the ratio for these earlier cohorts never rises to the level it does for surviving users from later cohorts.

Chapter 5

Communities Trajectories

Chapter 6

Conclusions and Final Remarks

6.1 Conclusion and Future Work escience15

This paper presented a novel method of capturing some of a user's non-explicit reactions to followees' content in Twitter by using text similarity scores between a user's tweets and those of their followees. The analysis indicates that the method does generate higher scores on average for system tagged Replies and Retweets than Non-Tagged tweets, suggesting that it captures real signal about responses. Using a conservative cutoff for predicting whether a non-tagged tweet is a response suggests that at least 11% of actual responses are not tagged by the system. These responses are distributed across almost a quarter of the users in the dataset, with a quarter of those having more missed reaction messages than explicit system tagged ones. These are not just naive, low-activity users who do not understand Twitter and might be ignored in analysis; a number of these users are quite active, with dozens or hundreds of tweets in a 14-day window.

Although the method has provided useful insights into the prevalence of non-explicit replies in Twitter, it is a coarse model. It tends to under-evaluate Replies; is more sensitive to Retweet size than desirable; likely misses a number of non-explicit responses that have lower scores but are nonetheless real responses to the feed; and doesn't address responses to content outside the feed such as views by hashtag or username. Ongoing work aims at addressing these limitations by improving the quality of the scoring function. One natural way of improving the scoring function is to incorporate other relevant social features highlighted by past work (Table 2.1). We expect that better models of language, network characteristics, and attention that build on these features would give better estimates of how people react to content produced by their followees.

Another possible unfolding research topic is how to use these reaction scores to understand the reaction patterns and estimate the individual reaction level for each user. This is important for effective models of diffusion at all levels, from understanding when adding an individual to a follower network might be most valuable, to estimating the overall reach of an individual's network, to modeling diffusion of information in the large. Missing 11% of responses and 6% users is a substantial amount of error to bear for such models, making the identification of non-explicit responses an important problem to pursue.

6.2 Discussion www16

In this section we discuss some of the processes that might explain our observations, and how they connect to other literature. We're not arguing here that we know the answers; instead, we see these as interesting avenues for future work.

6.2.1 Why are newer "active" users less so?

We have seen that users from later cohorts have a lower posting average than in earlier cohorts. One plausible explanation is that users self-select: users that find Reddit early in its life are also

more likely than average to be those who will be attracted to it. Previous work has shown that online book reviews have a self-selection bias, where people who are more likely to like (or promote) the book review it earlier, leading to a positive early bias in an item's life [LH08]. In Reddit's case, this would mean that the mixture of users joining in the early stage of the community would be disproportionately likely to be the most active ones and the latter ones are more likely to be less active.

Another plausible hypothesis for later cohorts having a higher number of less active users could be that, over time, Reddit has accumulated an increasing number of valuable-but-small/niche communities. The increased diversity might support a wider set of users in getting value, explaining the increased survival percentage. The niche/smaller nature of newer communities might provide fewer opportunities to both submit and comment, explaining the lower average activity for surviving users.

A third hypothesis is that Reddit overall is becoming more about consumption and voting on content rather than producing it. Older users with contribution norms continue to contribute; newer users tend to provide audiences and feedback. High-resolution voting data could be a real boon in understanding if this is true.

6.2.2 Why are comments getting shorter?

We also observed that overall, comment lengths are getting shorter over time.

One hypothesis is that users are being shaped by an "initial value problem". We can imagine that users as they join the network, tend to produce content according to the norms of what they see [KGM10, DNMWJP13]. The observed behavior of the comments length for the users in Reddit is a initial drop, followed by steady increase as the user survives. If the starting point for the initial drops are taken as the average of the network, that is what is observed by the user in the network, the initial drop would place each cohort starting at lower levels than the previous one. Figure 4.6a presents some support this hypothesis: the initial month of each cohort year, which consists of data only from users who joined in that month, is quite close to the overall line from the prior month.

Another hypothesis advanced by community members[1] is that reddit's karma system favors shorter comments. That is, people can get more upvotes for a given amount of effort by writing more, shorter comments. This could be directly measured even with the available data, and might be the start of a very interesting line of future work that tries to model strategic posting and attention distribution behavior in Reddit.

6.2.3 Why do comments per submission increase?

We also saw that comments per submission increase over time for surviving users, and that this is most dramatic for users who join earlier.

One process hypothesis is that this is because early in Reddit's life, there simply weren't as many submissions to comment on, meaning that people who wanted to be active contributors more or less had to submit in order to do so. As the community grew, more content became available, making information seeking more valuable—and perhaps increasing the value and ease of commenting.

This question of ease and value might be more general, and tie to our earlier observations about self-selection and karma accumulation. Most users in social networks are known to be lurkers: users that only seek information and passively observe, not engaging and contributing to the network [RRS04, NP00]. Consumption is valuable and easy, and in reddit, some contributions are easier than others: reading is easier than voting; voting is easier than commenting; commenting is easier than submitting. Only users for whom finding and submitting comment is relatively easy or relatively valuable are likely to be frequent submitters or "power users" [PHT09, KCP+07]—and we suspect such users are more likely to be ones who found Reddit earlier on and stuck with it when it was relatively small.

[1] See https://www.reddit.com/r/TheoryOfReddit/comments/1a7aoj/retracing_the_evolution_of_reddit_through_post/.

6.2.4 Limitations and Future Work

In this paper we focused our attention on behavior attributable to specific users, which in this dataset meant submissions and comments. As with many analysis that focus on visible behavior, this means we miss important phenomena. In particular, we discount lurkers despite their known importance as audience members [NPF03] and potential future contributors [RGA06]. Many lurkers likely vote, and thus lurking may be even more important in a context like Reddit where votes affect content visibility and provide explicit markers of attention and reputation.

However, the dataset does not have information on individual voters or timestamps, just the aggregate number of votes a post had received at the time of the crawl, making it impossible to effectively treat them as activity measures and ways to understand the behavior of those who voted. The existing voting data might be much more useful, however, in addressing questions that involve predicting a given user's future behavior based on whether and how other users respond to a user's early contributions [JK06, SWL12]

Another blind spot that focusing on visible behavior can induce is our emphasis on active users. This is a reasonable view of the community that focuses on what is happening, but our results should all be interpreted in the context of "given the set of active users at any given time". Applying these results to questions that require considering all users would be a mistake.

We did, implicitly, consider survival in the analyses that broke cohort down by survival time; we see careful thinking about what it means to "survive" in a community as an interesting problem in its own right. Potentially, users' "breaks" from the network can influence both our results and other analyses that assume users depart on their last visible day of activity. Focusing on activity also fails to account for actual deletion in many contexts. In Reddit, activity from users is marked with a username of "[deleted]" (which we were able to ignore after realizing that one author had millions of comments!), but in some contexts, such as Wikipedia articles that are deleted, edit behavior on those articles do not show up in many data dumps.

These questions of how to define active users and dead users and distinguishing patterns of behavior seems an interesting venue to pursue. Better definitions of "active" and "dead" users might allow us to characterize the burstiness of their behavior. Some users might only interact with the network in some specific occasions while some users might have a much more uniform pattern; in Wikipedia, the practice of leaving temporarily is so common it is called a "wikibreak". Understanding how your network fares in terms of user burstiness is essential to understand how the users use the network and to shape the user experience. A better definition of "death" would allow us to investigate the "rebirth" of users, that is, users that come back to the community. Rather than an annoying right censorship statistical problem, it might pose a much more central issue, as a community's survival might not depend only in its ability to attract and retain users, but also in the ability to "resurrect" old users.

6.3 Conclusions

This work highlights the importance of taking time into consideration when analyzing users' evolution in social networks. We do so by cohorting the users based on their creation year. Although simple, this approach provides evidence of significant differences between methods that account for time with methods that do simple overall analyses. We also analyze the evolution of users and communities from a shifted time referential: considering the time of an action in relation to the user creation date. This also reveals unexpected phenomena that we would otherwise not notice.

From the user perspective, we found that user posting activity for surviving Reddit users is actually significantly higher than a naive average would suggest, that older users who survive are considerably more active than younger survivors, and that these newer users are unlikely to catch up. Controlling for survival, we also found that users have a stable level of posting activity over time (with slightly decreasing patterns) and the percentage of surviving but low-activity users is increasing in the younger cohorts.

Similarly, we analyzed user effort based on average comment length. We found that, while the overall average in Reddit seems to decrease, users actually write longer comments as they survive, no matter when they joined. Still, later cohorts of users that joined the network are writing smaller comments; their greater number leads to this version of Simpson's paradox, where where the overall average decreases while the series for each individual cohort increases.

Finally, we analyze the type of activities users engage, differentiating comments and submissions. We found that users with a higher comments per submission ratio are more likely to survive longer in the network. More than that, this behavior changes as the users survive—particularly for the early cohorts. Users change their comments per submissions patterns, and their main mechanism to do so seems to be replacing their submitting by commenting behavior, since their posting activity remains stable. We also discussed a possible explanation for this observation based on commenting being partially an information seeking task with an associated effort between lurking and submitting. This made it less likely in early Reddit, in which content was not as available as in later years.

Both our and work and its limitations suggest fruitful directions for better understanding of users' evolution in both Reddit and online communities in general, directions we hope inspire other work in this area.

Appendix A

Tweets' Scores

Table 1: *Pairs of users' tweets (top in each row) and highest scoring messages in the windows (bottom in each row) for Retweets, Replies, and Non-Tagged tweets. Tweets were randomly selected across the range of scores in each set.*

#	Score	Retweets
1	1.0	**brandonlondon:** RT @neiltyson: A H R B Q D W E F L M N S X G I J K O P C T V Y U Z - Gotta love what the alphabet looks like in alphabetical order.
		neiltyson: A H R B Q D W E F L M N S X G I J K O P C T V Y U Z - Gotta love what the alphabet looks like in alphabetical order.
2	0.834	**michael_palko:** RT @8_Semesters: A girlfriend would be great, but I'm already in a pretty committed relationship with alcoholism and bad decisions.
		8_Semesters: A girlfriend would be great, but I'm already in a pretty committed relationship with alcoholism and bad decisions.
3	0.768	**mike_sprague:** RT @mshowalter: All the weird horny stuff between Glenn and Maggie on Walking Dead makes me very uncomfortable.
		mshowalter: All the weird horny stuff between Glenn and Maggie on Walking Dead makes me very uncomfortable.
4	0.602	**__ShesBrownSKIN:** RT @_CarGotThat: Brandywine Came Out With Win, #TeamBwine
		_CarGotThat: Brandywine Came Out With Win, #TeamBwine
5	0.522	**Becchappell:** RT @oliviaaajayne_: I just love Toy Story, all of them
		oliviaaajayne_: I just love Toy Story, all of them
6	0.408	**__tiki:** RT @Greektown1921: Welp now you know
		Greektown1921: Welp now you know
7	0.303	**terrigolas:** RT @jam_bu88: Facebook is down? Oh no, how are cancer and child abuse going to stop without all those likes? :(
		dsilverman: RT @CalebGarling: Stop acting like we have 'rights' on Facebook http://t.co/geE9NjHH
8	0.248	**Becchappell:** RT @j4kebro: going to be slightly awkward when Jahmene scans the xfactor winner's single in Asda
		bombaytricycle: RT @justaholyfooool: Jahmene is gonna be scanning James Arthers CD at ASDA now.. awks
9	0.132	**HollywoodLadyj:** RT @RealSkipBayliss: RG3 should give Michael Vick a class in scrambling.
		ESPN_FirstTake: RG3 running at 4G!
10	0.070	**davidAmejia:** RT @Snoopy: It's Monday, Snoopy! http://t.co/asOF9yPA
		AshKetchum151: Mondays are like Zubats. Nobody likes Zubats.

#	Score	Replies
11	0.768	**esterrick:** RT @StationBistro1: On deck - Next week's soup is White Bean and Smoked Turkey Chili!
		StationBistro1: On deck - Next week's soup is White Bean and Smoked Turkey Chili!
12	0.693	**Serrae:** @MollytheGhost @PhantomRat @hollye83 @hockeybychoice @onlymystory @sjopierce @phouse1964 Hate them.
		hollye83: @hockeybychoice @onlymystory @PhantomRat @sjopierce @phouse1964 @MollytheGhost @Serrae Hateful. Just hateful.
13	0.573	**HectorBesmonte:** @EmmittWard @jccassiel @hottiemarkie33 @mark_purdie @Rhino108 @JasonReedyOH420 yeah! thanks emmit! muah! love, Hugs! for you!
		EmmittWard: @HectorBesmonte @jccassiel @hottiemarkie33 @mark_purdie @Rhino108 @JasonReedyOH420 happy birthday hector sending you love and hugs buddy.
14	0.477	**VinnyG5:** @shanmilanowski I wasn't really that drunk that day...I wouldn't get hammered and let you drive with me...but that's a secret so shhhhh
		shanmilanowski: RT @VinnyG5: @shanmilanowski lets do that thing were we get drunk and drive around the city while I'm playing my guitar ......again.
15	0.386	**RazWorth:** @SophieRaby yeah! Buzzin
		SophieRaby: @RazWorth do you? :(
16	0.283	**SarahMcCallumXX:** @SarahMcCallumXX @mton1996 forgot the x hahaa
		mton1996: @SarahMcCallumXX aw hen, I feel for you x
17	0.245	**missRaichl:** @michel_andness good morning!
		Michel_andNess: Good morning everyone.
18	0.168	**Mahalia_Enares:** @kiafranklins_ HAPPY BIRTHDAY!:) xx
		istoleursmartie: @Mahalia_Enares i can be!!
19	0.133	**MeganDoesNOLA:** @KurlyKonfektion Niiiiice...
		KurlyKonfektion: @MeganDoesNOLA lmao! I'm gonna put a slice of bacon with/in my drink and see what happens lol
20	0.068	**essfardella:** @Ali_Diesel_ There it is.
		Ali_Diesel_: RT @shkeeber: I am not a slut. I'm an erection enthusiast.

#	Score	Non-Tagged
21	0.920	**hypervocal:** RT @Reuters: FLASH: #Egypt's Mursi has left presidential palace, two presidency sources say after protesters, police clash outside.
		AntDeRosa: RT @Reuters: FLASH: #Egypt's Mursi has left presidential palace, two presidency sources say after protesters, police clash outside.
22	0.884	**hypervocal:** It's time. RT @whitehouse: hey guys - this is barack. ready to answer your questions on fiscal cliff & #my2k. Let's get started. -bo
		ethanklapper: RT @whitehouse: hey guys - this is barack. ready to answer your questions on fiscal cliff & #my2k. Let's get started. -bo
23	0.782	**Mrjscott:** A white woman... RT @T_dot_Lee: A woman? RT @majic1021: 'Fresh Prince' Star Alfonso Ribeiro Weds http://t.co/IfT3Zqlr
		T_dot_Lee: A woman? RT @majic1021: 'Fresh Prince' Star Alfonso Ribeiro Weds http://t.co/iTrfZfeM
24	0.697	**wulan_kyuufilan:** RT @WestlifeFansite: hear @nickybyrneoffic on the radio one minute ago!! it was funny :D x
		WestlifeFansite: hear @nickybyrneoffic on the radio one minute ago!! it was funny :D x
25	0.579	**esterrick:** RIP Mr Brubeck. Take five. "@annesaurus: Dave Brubeck, jazz icon, dead at 91. http://t.co/ae9UIRmP"
		Supperphilly: RT @annesaurus: Dave Brubeck, jazz icon, dead at 91. http://t.co/sOrBOFBR
26	0.443	**Zac_Hartlage14:** @BadJerry20 OKC traded James Harden
		24_Jag: Why WOULD OKC TRADE JAMES HARDEN????
27	0.359	**Serrae:** (that should have had a link to the Tina and Amy host Golden Globes article. But I'm too lazy to fix it now)
		MichaelAusiello: Genius Move: Tina Fey and Amy Poehler to Host 2012 Golden Globe Awards! http://t.co/zdc2hS8F
28	0.275	**nicoleoraha:** @Niallofficial are you excited to come to Australia and meet all your amazing fans like me? ;) xx #asknialler 11
		ahoynialler: @NiallOfficial EXCITED TO COME BACK TO AUSTRALIA, cause we miss you lots xox #asknialler
29	0.242	**Serrae:** All of @fatherdowling's #captainhottie pirate puns for #ouat are perfect. It's the reason we are twitter friends.
		fatherdowling: I apologize in advance for inappropriate pirate puns. #sorryImnotsorry #OUAT
30	0.139	**ESTL63:** Forever my lady lol
		DaMontesMom_415: @VictoriaLMathis just go get it!! Lol (the devil) aint I? Lol but I would lol

Bibliography

[ABJ06] Jaime Arguello, Bs Butler e Elisabeth Joyce. Talk to me: foundations for successful individual-group interactions in online communities. *Proceedings of the . . .*, páginas 959–968, 2006. 21

[AHP12] Karl L. Alexander, Scott Holupka e Aaron M. Pallas. Social Background and Academic Determinants of Two-Year versus Four-Year College Attendance : Evidence from Two Cohorts a Decade Apart. *American Journal of Education*, 96(1):56–80, 2012. 6

[APG12] Yoav Artzi, Patrick Pantel e Michael Gamon. Predicting responses to microblog posts. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012. 2, 3, 5

[AW93] Orazio P Attanasio e Guglielmo Weber. Consumption Growth, the Interest Rate and Aggregation. *Review of Economic Studies*, 60(3):631–49, 1993. 6

[BBF+05] S Bryant, S Bryant, A Forte, A Forte, A Buckman e A Buckman. Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work. Em *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, páginas 1–10. ACM, 2005. 6

[Bel05] S. Beldona. Cohort Analysis of Online Travel Information Search Behavior: 1995-2000. *Journal of Travel Research*, 44(November):135–142, 2005. 6, 7

[BF15] Shlomo Berkovsky e Jill Freyne. Personalised Network Activity Feeds: Finding Needles in the Haystacks. Em *Mining, Modeling, and Recommending 'Things' in Social Media*, volume 8940, páginas 21–34. Springer International Publishing, 2015. 4

[BJP08] Brian Butler, Elisabeth Joyce e Jacqueline Pike. Don't Look Now, But We've Created a Bureaucracy : The Nature and Roles of Policies and Rules in Wikipedia. *CHI 2008 Proceedings*, páginas 1101–1110, 2008. 4

[BKL09] Steven Bird, Ewan Klein e Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edição, 2009. 10

[BMAC13] Matthew Burgess, Alessandra Mazzia, Eytan Adar e Michael Cafarella. Leveraging Noisy Lists for Social Feed Ranking. *Association for the Advancement of Artificial Intelligence*, 2013. 4

[BRMA12] Eytan Bakshy, Itamar Rosenn, Cameron Marlow e Lada Adamic. The role of social networks in information diffusion. *WWW 2012 – Session: Information Diffusion in Social Networks April 16–20, 2012, Lyon, France*, páginas 519–528, 2012. 1

[BWCD11] L. W. Black, H. T. Welser, D. Cosley e J. M. DeGroot. Self-Governance Through Group Discussion in Wikipedia: Measuring Deliberation in Online Groups. *Small Group Research*, 42:595–634, 2011. 1

[CAKL10] Boreum Choi, Kira Alexander, Robert E Kraut e John M Levine. Socialization Tactics in Wikipedia and Their Effects. *Cscw*, páginas 107–116, 2010. 6, 19

[CC12] Giovanni Comarela e Mark Crovella. Understanding factors that affect response rates in twitter. *HT '12 Proceedings of the 23rd ACM conference on Hypertext and social media*, 2012. xv, 1, 3, 5, 9, 14

[CHdZ10] Teresa Correa, Amber Willard Hinsley e Homero Gil de Zúñiga. Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, 26(2):247–253, 2010. 2, 7

[CHK10] Dan Cosley, Daniel Huttenlocher e Jon Kleinberg. Sequential influence models in social networks. *... on Weblogs and Social ...*, páginas 26–33, 2010. 2

[DAE+10] Gary L. Davis, Miriam J. Alter, Hashem El–Serag, Thierry Poynard e Linda W. Jennings. Aging of Hepatitis C Virus (HCV)-Infected Persons in the United States: A Multiple Cohort Model of HCV Prevalence and Disease Progression. *Gastroenterology*, 138(2):513–521.e6, 2010. 6

[DNMWJP13] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky e Christopher Potts. No Country for Old Members : User Lifecycle and Linguistic Change in Online Communities. *Proceedings of the 22nd international conference on World Wide Web*, páginas 307–317, 2013. 2, 4, 6, 7, 32

[DR01] Pedro Domingos e Matt Richardson. Mining the Network Value of Customers. *Proceedings of the Seventh {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining*, páginas 57–66, 2001. 1

[FD88] Glenn Firebaugh e Kenneth E. Davis. Trends in Antiblack Prejudice, 1972-1984: Region and Cohort Effects. *American Journal of Sociology*, 94(2):251, 1988. 6

[GBL10] Amit Goyal, Francesco Bonchi e Laks V.S. Lakshmanan. Learning influence probabilities in social networks. *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, página 241, 2010. 4

[GK09] Eric Gilbert e Karrie Karahalios. Predicting tie strength with social media. *ACM Conference on Human Factors in Computing Systems*, páginas 211–220, 2009. 1, 4

[Gle05] Norval D Glenn. *Cohort analysis*, volume 5. Sage, 2005. 6

[GLNGT04] D. Gruhl, David Liben-Nowell, R. Guha e A. Tomkins. Information diffusion through blogspace. *ACM SIGKDD Explorations Newsletter*, 6(2):43–52, 2004. 19, 22

[Goy08] Kimberly A. Goyette. College for some to college for all: Social background, occupational expectations, and educational expectations over time. *Social Science Research*, 37(2):461–484, 2008. 6

[GTC+09] Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang e Yihong (Eric) Zhao. Analyzing Patterns of User Content Generation in Online Social Networks. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, páginas 369–378, 2009. 19, 22

[HDD11] Liangjie Hong, Ovidiu Dan e BD Davison. Predicting popular messages in twitter. *WWW '11 Proceedings of the 20th international conference companion on World wide web*, página 57, 2011. 3, 5

[HECR96] Ralph I Horwitz, Levy EM, Viscoli CM e Horwitz RI. The effect of acute renal failure on mortality: A cohort analysis. *JAMA*, 275(19):1489–1494, 1996. 6

[HMS11] John Hannon, Kevin McCarthy e Barry Smyth. Finding Useful Users on Twitter: Twittomender the Followee Recommender. *Springer-Verlag Berlin Heidelberg*, 6611:784–787, 2011. 4

[HP09] Amanda Lee Hughes e Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(May):248, 2009. 19, 21

[Ing14] Mathew Ingram. Digg Redesign Met With a Thumbs Down, 2014. 6

[Jen96] M. Kent Jennings. Political Knowledge Over Time and Across Generations. *Public Opinion Quarterly*, 60:228–252, 1996. 6

[JK06] Elisabeth Joyce e Robert E. Kraut. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11:723–747, 2006. 1, 33

[JSFT07] Akshay Java, Xiaodan Song, Tim Finin e Belle Tseng. Why We Twitter: Understanding Microblogging. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, páginas 56–65, 2007. 19, 21

[KCP+07] a Kittur, E Chi, B a Pendleton, B Suh e T Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *Algorithmica*, 1(2):1–9, 2007. 32

[KGM10] Farshad Kooti, Krishna P Gummadi e Winter a Mason. The Emergence of Conventions in Online Social Networks. *Artificial Intelligence*, páginas 194–201, 2010. 2, 4, 6, 32

[KSPC07] Aniket Kittur, Bongwon Suh, Bryan A Pendleton e Ed H Chi. He Says, She Says: Conflict and Coordination in Wikipedia. Em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, number February 2011 in CHI '07, páginas 453–462, New York, NY, USA, 2007. ACM. 6

[Lar14] F Lardinois. Digg redesign tanks: Traffic down 26%(updated with new reddit stats), 2014. 6

[LH08] Xinxin Li e Lorin M. Hitt. Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474, 2008. 6, 32

[LJ13] Zhe Liu e BJ Jansen. Factors influencing the response rate in social question and answering behavior. *CSCW '13 Proceedings of the 2013 conference on Computer supported cooperative work*, página 1263, 2013. 3, 5

[LKC08] Kevin Lewis, Jason Kaufman e Nicholas Christakis. The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network. *Journal of Computer-Mediated Communication*, 14(1):79–100, 2008. 2

[LKF05] Jurij Leskovec, Jon Kleinberg e Christos Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. *In KDD*, páginas 177–187, 2005. 4

[LOTW13] Zhunchen Luo, Miles Osborne, Jintao Tang e Ting Wang. Who Will Retweet Me? Finding Retweeters in Twitter. Em *Proceedings of the 19th International Conference on World Wide Web*, páginas 5–8, 2013. 3, 5

[LW84] Mark R Levy e Sven Windahl. Audience activity and gratifications a conceptual clarification and exploration. *Communication research*, 11(1):51–78, 1984. 19

[MCT15] Hannah J. Miller, Shuo Chang e Loren G. Terveen. "I LOVE THIS SITE!" vs. "It's a little girly": Perceptions of and Initial User Experience with Pinterest. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, páginas 1728–1740, 2015. 2, 4

[MF12] William M Mason e Stephen Fienberg. *Cohort analysis in social research: Beyond the identification problem*. Springer Science & Business Media, 2012. 6

[ML13] J McAuley e J Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. *Proceedings of the 7th ACM conference on Recommender systems - RecSys '13*, páginas 165–172, 2013. 4

[NP00] Blair Nonnecke e Jenny Preece. Lurker demographics: Counting the silent. *Proceedings of the SIGCHI conference on ...*, 2(1):1–8, 2000. 32

[NPF03] Blair Nonnecke, Jenny Preece e Danyel Fisher. Silent participants: Getting to know lurkers better. Em *From usenet to CoWebs*, páginas 110–132. Springer, 2003. 33

[PCL+07] Reid Priedhorsky, Jilin Chen, Shyong Tony K Lam, Katherine Panciera, Loren Terveen e John Riedl. Creating, destroying, and restoring value in wikipedia. *Proceedings of the 2007 international ACM conference on Conference on supporting group work - GROUP '07*, página 259, 2007. 2

[PES10] Jm Pujol, Vijay Erramilli e Georgos Siganos. The little engine (s) that could: scaling online social networks. *Acm Sigcomm'10*, páginas 375–386, 2010. 2

[PGC+10] Tim Paek, Michael Gamon, Scott Counts, David Maxwell Chickering e Aman Dhesi. Predicting the Importance of Newsfeed Posts and Social Network Friends. *Artificial Intelligence*, páginas 1419–1424, 2010. 4

[PHT09] Katherine Panciera, Aaron Halfaker e Loren Terveen. Wikipedians are born, not made: a study of power editors on Wikipedia. *Human Factors*, páginas 51–60, 2009. 2, 4, 19, 23, 32

[POL11] S Petrovic, Miles Osborne e Victor Lavrenko. RT to Win! Predicting Message Propagation in Twitter. *ICWSM '11 International AAAI Conference on Weblogs and Social Media*, 2011. 3, 5

[PPET10] Katherine Panciera, Reid Priedhorsky, Thomas Erickson e Loren Terveen. Lurking? cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki. *CHI '10 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, páginas 1917–1926, 2010. 4

[Pre01] Marc Prensky. Digital Natives, Digital Immigrants - Prensky - Digital Natives, Digital Immigrants - Part1.pdf. *MCB University Press, Vol. 9 No. 5*, páginas 1–6, 2001. 7

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 11

[PZP11] HK Peng, Jiang Zhu e Dongzhen Piao. Retweet modeling using conditional random fields. *ICDMW '11: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, páginas 336–343, dec 2011. 3, 5

[RGA06] C Ridings, David. Gefen e Bay Arinze. Psychological Barrier: lurker and poster motivation and behaviour in online communities . *Communication of Association for Information Systems*, 18:329–354, 2006. 33

[RRS04] S. Rafaeli, G. Ravid e V. Soroka. De-lurking in virtual communities: a social communication network approach to measuring the effects of social and cultural capital. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004*, 00(C):1–10, 2004. 32

[Ryd65] Norman B. Ryder. The cohort as a concept in the study of social change. *Americal Sociological Review*, 30(6):843–861, 1965. 7

[SGC13] A Sharma, M Gemici e Dan Cosley. Friends, Strangers, and the Value of Ego Networks for Recommendation. *ICWSM*, 2013. 12

[SHPC10] Bongwon Suh, Lichan Hong, Peter Pirolli e EH Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. *2010 IEEE Second International Conference on Social Computing (SocialCom)*, páginas 177–184, aug 2010. 1, 3, 5

[SM11] Salvatore Scellato e Cecilia Mascolo. Measuring user activity on an online location-based social network. *2011 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2011*, páginas 918–923, 2011. 19, 21

[SOS12] Markus Schaal, J O'Donovan e Barry Smyth. An analysis of topical proximity in the twitter social graph. *Social Informatics*, páginas 232–245, 2012. 4

[SS51] Royal Statistical Society e E H Simpson. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):pp. 238–241, 1951. 26

[SWL12] Chandan Sarkar, Donghee Yvette Wohn e Cliff Lampe. Predicting length of membership in online community "everything2" using feedback. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion - CSCW '12*, página 207, 2012. 33

[TL15] Chenhao Tan e Lillian Lee. All Who Wander : On the Prevalence and Characteristics of Multi-community Engagement. *WWW '15*, abs/1503.0, 2015. 4

[Tom08] Jure LeskovecandLars BackstromandRavi KumarandAndrew S Tomkins. Microscopic evolution of social networks. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, páginas 462–470, 2008. 22

[UC11] I Uysal e W B Croft. User oriented tweet ranking: A filtering approach to microblogs. *International Conference on Information and Knowledge Management, Proceedings*, páginas 2261–2264, 2011. 4

[WCK+11] Howards T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay e Marc Smith. Finding social roles in Wikipedia. *Proceedings of the 2011 iConference on - iConference '11*, páginas 122–129, 2011. 2, 6, 12, 19, 28

[WGF07] Howard T. Welser, Eric Gleave e Danyel Fisher. Visualizing the Signatures of Social Roles in Online Discussion Groups. *Journal of Social Structure*, 8(2):1–32, 2007. 28

[WLZL12] Xufei Wang, Huan Liu, Peng Zhang e Baoxin Li. Identifying Information Spreaders in Twitter Follower Networks. Relatório técnico, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, 2012. 3, 5

[YC09] Jiang Yang e Scott Counts. Predicting the Speed , Scale , and Range of Information Diffusion in Twitter. *Fourth International AAAI Conference on Weblogs and Social Media*, páginas 355–358, 2009. 4

[ZKK14] Haiyi Zhu, Robert E Kraut e Aniket Kittur. The impact of membership overlap on the survival of online communities. Em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, páginas 281–290. ACM, 2014. 2

Index