# Revealing social networks' missed behavior: detecting reactions and time-aware analyses

Samuel Martins Barbosa Neto

Thesis presented
to the
Institute of Mathematics and Statistics
of the
University of São Paulo
as requirement
to obtain the title
of
Doctor of Philosophy

Program: Computer Science

Advisor: Prof. Dr. Roberto Marcondes Cesar Jr.

Coadvisor: Dr. Claudio Santos Pinhanez

Foreign Advisor: Prof. Dr. Dan Cosley

The author received financial support from CAPES/CNPq/FAPESP

São Paulo, July 2017

**Revealing social networks' missed behavior: detecting reactions and time-aware analyses**

This version of the thesis contains the corrections and suggestions given by the Judging Committee during the defense of the original version of the work, that took place on May 29th, 2017. A copy of the original version is available at the Institute of Mathematics and Statistics of the University of São Paulo.

Judging Committee:

- Prof. Dr. Roberto Marcondes Cesar Junior — IME — USP
- Dr. Claudio Santos Pinhanez — IBM Research Brazil
- Prof. Dr. Denis Deratani Mauá — IME — USP
- Prof. Dr. Jesus Mena—Chalco — UFABC
- Prof. Dr. Luciano Digiampietri — EACH — USP

# Acknowledgments

I am greatly thankful to the effort and support of Professor Dan Cosley, who contributed to this thesis in uncountable ways. This was an enjoyable 6-year journey, which I dedicate to family, friends and many others who provided the most varied contributions. Some of these people are still around, others drifted away, but I will forever cherish the memories and lessons I have learned from every single one of them.

# Resumo

Barbosa Neto, S. M. **Revelando o comportamento perdido em redes sociais: detectando reações e análises temporais**. 2017. 68f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

Comunidades online proporcionam um ambiente fértil para análise do comportamento de indivíduos e processos sociais. Por exemplo, ao modelarmos interações sociais online, é importante compreendemos quando indivíduos estão reagindo a outros indivíduos. Além disso, pessoas e comunidades mudam com o passar do tempo, e levar em consideração sua evolução temporal nos leva a resultados mais precisos. Entretanto, em muitos casos, o comportamento dos usuários pode ser perdido: suas reações ao conteúdo ao qual são expostos não são capturadas por indicadores explícitos (*likes* no Facebook, *replies* no Twitter). Agregações temporais de dados pouco criteriosas podem ocultar, enviesar ou até levar a conclusões equivocadas sobre como usuários evoluem.

Apresentamos uma nova abordagem para o problema de detectar respostas não-explicitas que utiliza similaridade tf-idf entre tweets de um usuário e tweets recentes que este usuário recebeu de quem segue. Com base em dados de postagens de um mês para 449 redes egocêntricas do Twitter, este método evidencia que temos um volume de ao menos 11% de reações não capturadas pelos mecanismos explicitos de *reply* e *retweet*. Além disso, essas reações não capturadas não estão uniformemente distribuídas entre os usuários: alguns usuários que criam *replies* e *retweets* sem utilizar os mecanismos formais da interface são muito mais responsivos a quem eles seguem do que aparentam. Isso sugere que detectar respostas não-explicitas é importante para mitigar viéses e construir modelos mais precisos a fim de estudar interações sociais e difusão de informação.

Abordamos o problema de evolução de usuários no Reddit com base em dados entre o período de 2007 a 2014. Utilizando métodos simples de diferenciação temporal dos usuários – cohorts anuais – encontramos amplas diferenças entre o comportamento, que incluem criação de comentários, métricas de esforço e sobrevivência. Desconsiderar a evolução temporal pode levar a equívocos a repeito de fenômenos importantes. Por exemplo, o tamanho médio dos comentários na rede decresce ao longo de qualquer intervalo de tempo, mas este tamanho é crescente em cada uma das cohorts de usuários no mesmo período, salvo de uma queda inicial. Esta é uma observação do Paradoxo de Simpson. Dividir as cohorts de usuários em sub-cohorts baseadas em anos de sobrevivência na rede nos fornece uma perspectiva melhor; usuários que sobrevivem por mais tempo apresentam um maior nível de atividade inicial, com comentários mais curtos do que aqueles que sobrevivem menos. Com isto, compreendemos melhor como usuários evoluem no Reddit e levantamos uma série de questões a respeito de futuros desdobramentos do estudo de comportamento online.

**Palavras-chave:** comportamento de usuário, rede social, paradoxo de simpson, twitter, reddit.

# Abstract

Barbosa Neto, S. M. **Revealing social networks' missed behavior: detecting reactions and time-aware analyses**. 2017. 68p. Thesis (Ph.D.) - Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2017.

Online communities provide a fertile ground for analyzing people's behavior and improving our understanding of social processes. For instance, when modeling social interaction online, it is important to understand when people are reacting to each other. Also, since both people and communities change over time, we argue that analyses of online communities that take time into account will lead to deeper and more accurate results. In many cases, however, users' behavior can be easily missed: users react to content in many more ways than observed by explicit indicators (such as likes on Facebook or replies on Twitter) and poorly aggregated temporal data might hide, misrepresent and even lead to wrong conclusions about how users are evolving.

In order to address the problem of detecting non-explicit responses, we present a new approach that uses tf-idf similarity between a user's own tweets and recent tweets by people they follow. Based on a month's worth of posting data from 449 ego networks in Twitter, this method demonstrates that it is likely that at least 11% of reactions are not captured by the explicit reply and retweet mechanisms. Further, these uncaptured reactions are not evenly distributed between users: some users, who create replies and retweets without using the official interface mechanisms, are much more responsive to followees than they appear. This suggests that detecting non-explicit responses is an important consideration in mitigating biases and building more accurate models when using these markers to study social interaction and information diffusion.

We also address the problem of users evolution in Reddit based on comment and submission data from 2007 to 2014. Even using one of the simplest temporal differences between users—yearly cohorts—we find wide differences in people's behavior, including comment activity, effort, and survival. Furthermore, not accounting for time can lead us to misinterpret important phenomena. For instance, we observe that average comment length decreases over any fixed period of time, but comment length in each cohort of users steadily increases during the same period after an abrupt initial drop, an example of Simpson's Paradox. Dividing cohorts into sub-cohorts based on the survival time in the community provides further insights; in particular, longer-lived users start at a higher activity level and make more and shorter comments than those who leave earlier. These findings both give more insight into user evolution in Reddit in particular, and raise a number of interesting questions around studying online behavior going forward.

**Keywords:** user behavior, social network, simpson's paradox, twitter, reddit.

# Contents

# List of Abbreviations

| | |
|---|---|
| tf | Term frequency |
| idf | Inverse document frequency |
| LDA | Latent Dirichlet Allocation |
| RT | Retweet |
| NLTK | Natural Language Toolkit |

# List of Symbols

| | |
|---|---|
| $u$ | User that is the center of the ego-network |
| $u_i$ | $i$-th user that is the center of an ego-network |
| $t_i$ | Tweet $i$ from user $u$ |
| $f_i$ | Followee $i$ |
| $ft_i$ | Followee's tweet $i$ |
| $w_i$ | Window composed of the $n$ most recent followees' tweets ($n = 100$ in our analysis) prior to tweet $t_i$ |
| $p_i^T$ | Percentage of the Tagged messages for a user $u_i$ in relation to the total of messages this user authored |
| $p_i^N$ | Percentage of the Non-Tagged messages for a user $u_i$ in relation to the total of messages this user authored |

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 A Brief History of this Thesis

Back in 2011 I was fresh out of college and, having majored in Applied Maths and Computer Engineering, I felt ready to face a bigger challenge. I applied for my Ph.D. at the University of São Paulo after talking to Professor Roberto. By the time, we discussed a few possibilities for a thesis subject and I made clear that I wanted something industry-related. Professor Roberto then suggested that I could work with Dr. Claudio Pinhanez, a former professor at the Maths Department of the University of São Paulo, that by then was a researcher at IBM Research Brazil. I was more than pleased with this plan.

After about one year into the program, I started interning at IBM Research Brazil. It was then that this thesis started to take form. By the time, Facebook was already the major online social network and the idea that modeling users' online behavior would be essential in the near future for the majority of businesses was something I had clear in my mind. Certainly, I was not the only one to think so, and there were many researchers already studying social networks such as Twitter, Wikipedia, LastFM, among others. Recommender systems were popular, as well as studies trying to predict virality of tweets.

My initial assignments as an intern were to craw data from the Twitter network that used for a study on the Obama's and Romney's electoral campaigns. Later on, this same data was used for the reaction detection work in this thesis. I also studied hidden Markov models in order to model users' behavior, thinking of the messages they authored as the observables for the model and their hidden states as their state of mind. Although this approach did not lead to any significant findings, it evolved into the idea of predicting weather or not a tweet message would be retweeted.

However, this idea of predicting retweets had been recently explored by the time. One thing though that was common in its related work by then was that none of them had a user-centered approach. They all gathered a lot of data and tried to find one classifier that would predict retweets or replies for any given message. As I was recently studying hidden Markov models, graphical models were fresh on my mind, and this lead to the proposal I made in my qualifying exam: create a graphical model for each user that could be used to classify his or her tweets and predict weather or not it would be retweeted/replied.

Soon after my qualifying exam, my internship at IBM ended and I started working at a startup named GetNinjas. At the same time, I was in contact with Professor Dan Cosley from Cornell University, to whom Dr. David Millen introduced me during my time at IBM. Professor Dan agreed to have me as a foreign student for one year, and six months after the end of my internship at IBM I was on my way to Cornell. During this six months, I focused on developing a geo-constrained distribution system for GetNinjas, that was an online services marketplace.

Arriving at Cornell, I started working at the Department of Information Science, together with Professor Dan Cosley. In retrospect, now I find that my thesis would be better contextualized in the Information Sciences department rather than the Computer Science, but I only realized this later on.

In the first months there, I worked on the research proposal I made during my qualifying exam. Unfortunately, the results of my proposed method were less than exciting: I often did not have enough data to properly train a bayesian network and when I had, sometimes my variable of interest, the retweet action, would be eliminated by a significance threshold. Discussing these issues with Dan, I dropped this approach. Soon after this I started working on a new concept, one that we had never seen anyone addressing before: sometimes users are reacting to things they have seen or been exposed to, but these reactions are not necessarily captured by the social networks explicit tools for reactions. We called these reactions implicit reactions.

Based on this idea, we proposed a method based on text similarity that would use ego-networks, a user-centered approach, to detect these implicit reactions that were missing when we addressed only retweets and replies. Since we were taking a user-centered approach, we were also able to measure how some users could be invisible if we only considered explicit reactions, and we saw that some users consistently avoided explicit reaction mechanisms, putting a lot more effort in copying and pasting rather than just clicking a "reply" or "retweet" button. These ideas were published in the 2015 e-Science conference and are part of this thesis.

Around the time I was working on the Twitter data, I learned of a large reddit dataset through another student at Cornell. Dan suggested that I explored the idea of novelty of content to predict the popularity of a new message. To do this, I time-ordered the messages in this reddit dataset and created time windows preceding these messages, trying to detect weather or not messages that had no similars in the network during this time window would be somehow more likely to propagate than messages that were more similar to content created during the time window. Unfortunately, we could not establish any correlation with novelty and the likelihood of propagating.

Moving forward, I came back to the observation I made during the Twitter work that users were varying the level effort they put in the network, and I thought that users could be optimizing their effort to reach some objective. I started trying to model users as a entities that have a bias towards some topics, and applied Latent Dirichlet Allocation to find "topics" over users. The idea was that a user could be seen as a document, and each time a user posted in a subreddit (similar to a forum in reddit, which is a collection of multiple subreddits) I could consider that this subreddit was "word" in the document. Based on this, I was able find topics of subreddits, clustering subreddits based on users' behavior. I wanted to understand how users would optimize their posting effort across the subreddits in these clusters.

While working with this idea, however, I faced many pitfalls while aggregating reddit's data (more than a billion comments), and realized that the behavior of users changed significantly over time. These pitfalls often hid important trends of growth or decrease in users' activity over time for the effort metrics I was trying to capture. We realized that these pitfalls were in fact occurrences of the Simpson's Paradox. We organized these observations and the interesting trends we discovered about how users evolved in reddit and published it in the 2016 WWW Conference. My favorite take away of this work is that the newer generations that join the social network are likely to become the majority of the network, and when looking at the average behavior of users, they will dominate the overall trend. This means that whenever you design products or strategies, you should always think more of who is going to join your network rather than who will stay in it, at least during the time it is growing exponentially.

By the time of the WWW conference, I was already back in Brazil. As soon as I arrived, GetNinjas got in touch with me and offered me a position as a Data Scientist, which I accepted. I began by working with their marketing system, which was primarily focused on AdWords by the time. This gave me a very clear idea of how to put all the things I had learned in practice: modeling users is a extremely powerful tool to optimize marketing strategies. I developed a bid optimization system and a campaign creation system for geo-constrained business. By the time of the defense of this thesis, the FAPESP research funding agency accepted our proposal to fund the development of this system through the PIPE program[1].

[1] This is a program to help develop research and new technology in small and medium business. More information can be found in http://www.fapesp.br/pipe/.

This summarizes the trajectory of how this thesis came to be and how the ideas presented here are connected in a larger context. I am greatly thankful for all the people that helped me and made all this possible.

## 1.2 Better Understanding Users' Behavior

Users' behavior span over a multitude of actions on social networks [BR09, GK09, CC12]: they search for friends, write messages, post images, videos and sounds among many other possibilities. Much of this content is captured by social networks mechanisms that explicitly tag and identify characteristics of the users' interaction. These mechanisms, however, are limited in capturing users' intention and diverse behavior. For instance, researchers have found that users might use references to content that only a specific audience from their peers can understand, turning a common context into a tool of privacy in public contexts [BM11]. The problem of identifying part of these interactions that are not fully captured is important to better understand users as well as to provide new perspectives on what kind of features social networks should provide. This is the main issue addressed in this thesis.

Research has shown that not accounting for time can lead to mistakes when dealing with social networks. Just as with offline contexts, systems and society changes over the years, as well younger generations have a different behavior from older generations. More than just considering different snapshots over time, we analyze users that joined at different stages of the network evolution. Just as with missed reactions, considering time-evolving users is important to reveal behavior that otherwise would not be noticed. Using simple cohorting methods, we demonstrate how different this behavior can be and how easy it is to draw wrong conclusions from averaging practices.

In the first part of this thesis, we address the problem of missed reactions, proposing a text similarity method to identify missed reactions and validating it on Twitter data. We show that a considerable number of users' reactions are not captured by current mechanisms in Twitter and that some types of users are significantly underrepresented based solely on these metrics. In the second part of this thesis, we analyze the users' evolution from a cohort perspective built on top of the time that they joined the network. We show, in the context of Reddit, how users' behavior vary depending on the year that they joined the network as well as how misleading not accounting for time can be.

This thesis was made in collaboration with IBM Research Brazil and Cornell University. The author interned at IBM Research for 2 years under the advisement of Claudio Pinhanez and visited Cornell University for 1 year as part of the sandwich Ph.D. program, under the advisement of Dan Cosley.

## 1.3 Objectives

The objective of this thesis is to improve our understanding of users' behavior on social networks. More specifically, to understand the missed behavior in the form of reactions that are not captured and temporal trends that are missed due to poor aggregation of data. We developed methods that help us to capture and identify missed reactions and propose analysis tools that allow us to avoid misunderstanding our data.

## 1.4 Contributions

We can summarize the contributions of this thesis as the following:

- Proposal of a new problem: understanding users' indirect reactions.
- Development of a method to detect indirect reactions in the context of Twitter, revealing about 11% of missed reactions, as well as patterns of behavior that are common, but not modeled, e.g., group conversations.

- A cohorted view of the users' evolution on Reddit over 7 years, revealing different behavioral patterns as a function of the users' tenure in the network.
- Practical examples of common aggregation practices that lead to wrong conclusions when dealing with time series.

The ideas proposed in this thesis have been published in [NGC+13, BCJC15, BCSC16].

## 1.5 Organization of this Thesis

We dedicate Chapter 2 of this thesis to the problem of detecting missed reaction. We propose a method based on text similarity to detect reactions other than retweets and replies in Twitter. We show that a significant amount of reactions are being missed. Furthermore, we show that many users consistently react in ways that are not captured by these mechanisms.

In Chapter 3, we cohort users in Reddit based on their creation and survival years. This analysis shows that naive aggregation of data can lead to wrong conclusions. It is shown that the proposed method can reveal users' evolution trends that would be otherwise missed. The proposed approach highlights the significant role of users joining and leaving the network in shaping the overall behavior.

Finally, in Chapter 4, we discuss and summarize our findings, also providing a discussion of their impacts and possible future venues of research to pursue.

# Chapter 2

# Using Text Similarity to Detect Social Interactions not Captured by Formal Reply Mechanisms[1]

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, one of the study cases of this thesis, 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 chapter assesses the prevalence of non-explicit responses in a dataset drawn from Twitter [NGC+13, BCJC15]. It is proposed 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[2] 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. Furthermore, 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

[1] The contents of this chapter were published at the 2015 IEEE 11th International Conference on e-Science (e-Science'15)[BCJC15]. Some adaptations were made to include further research and adjust the format to this thesis.

[2] We evaluate the average score for sets of messages that we know that are reactions. We end up picking the retweets' average, specially since they are a copy of the original message.

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. Later work [TFDN+16] studies a similar problem of identifying implicit responses from a provenance point of view. It uses a similar tf-idf method [DNTD+15], although considering a more holistic approach in contrast with a ego-network, user-centered approach[1]. We discuss the similarities and the results of both methods.

## 2.1 Related Work

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 thesis focus on the prevalence of implicit responses and 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

[1] The ego-network for the node $U$ can be defined as the sub-network of the neighbours of $u$ and $u$ itself, together with all the edges connecting these nodes. You can grown an ego-network by adding neighbours of neighbours, considering paths of length greater than 1 as part of the network as well.

| 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] |

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

and optimize exhibition of interesting content (messages or people, for instance) according to each user's particular preferences and context. In their survey, approaches to feed personalization are divided in 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.

Just as with retweet prediction, approaches that estimate tie strength and promote feed personalization rely on explicit interactions for their analyses. With that in mind, being able to identify non-explicit responses might improve these models.

Most related to this work 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 Twitter Dataset

To study potential responses of social network users we take an approach that considers the information users receive and emit. This 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 explored in this chapter 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 work 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. An overview of this dataset can be seen in Talbe 2.2.

### 2.2.1 IBM Internship

The author prepared this dataset during his internship at IBM Research Brazil with the co-advisorship of Dr. Claudio Pinhanez. During the extent of the internship, this dataset served as basis for other works in collaboration with IBM's researchers on the fields of agent-based simulations and information diffusion. More specifically, we built a system that would take the network structure

as an oriented graph where users are nodes and edges are "following" relationships, as well as the messages that were sent though these edges. We then performed a sentiment classification on these messages regarding two topics: Barack Obama and Mitt Romney. Based on each node sentiment emission distribution over the topics and a message's forwarding probability given it was received by one of its peers, we simulated the diffusion and prevalence of such messages on the network. We estimated how the volume of such messages would evolve over time for the whole network, as well as the proportions of positive and negative messages for each of the topics [GAdS+13, GAP+13, NGC+13].

While the experiments showed promising venues of how simulation approaches help us to understand diffusion and sentiment propagation in a network, they do not consider implicit reactions of users and non-explicit behavior, subjects that are the focus of this thesis. Also, while the author participated actively in the crawling and initial analysis, most of the simulation approach and models used were contributions of the IBM's researchers. Therefore we chose to highlight these as part of the contribution to the scientific community as an unfolding of the author's Ph.D. trajectory, but not include their full content on the thesis.

### 2.2.2 Crawler's technical aspects

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].

## 2.3 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

[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.

| | |
|---|---|
| Users | 449 |
| Tweets | 26051 |
| Average Tweets/User | 58.02 |
| Min Tweets/User | 1 |
| Max Tweets/User | 832 |
| Retweets | 5209 |
| Replies | 4192 |

**Table 2.2:** *An overview of the data collected for this work.*

$u$'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 chapter's first research question (**Ch.2 RQ1**), 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 (**Ch.2 RQ2**) 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?

### 2.3.1 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] studies 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 2.1 illustrates the window.

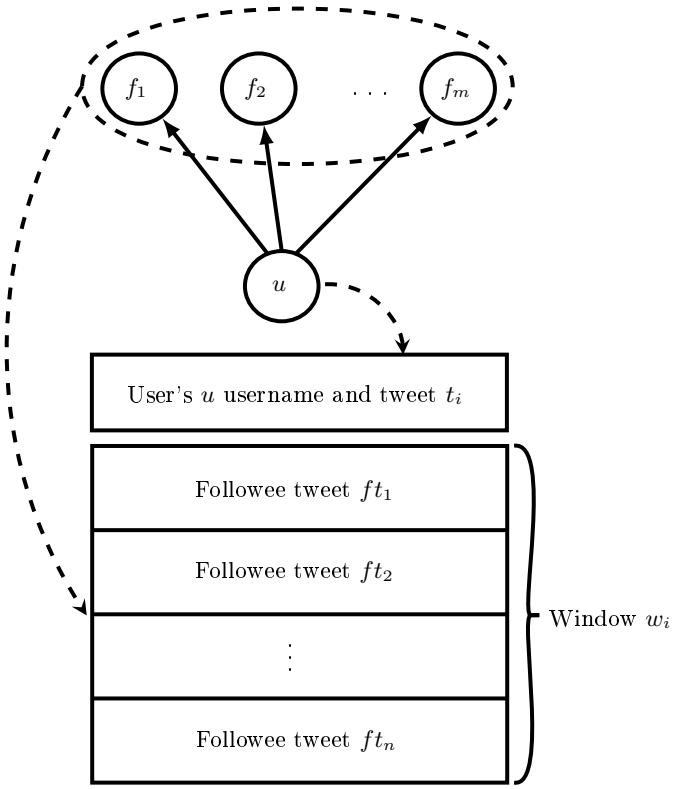



**Figure 2.1:** *Construction of the window $w_i$ for the tweet $t_i$. The tweets in the window (in this work, $n = 100$) are those most generated by user $u$'s followees most recently before $t_i$.*

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

**Table 2.3:** *Regular expressions used to extract features from tweets.*

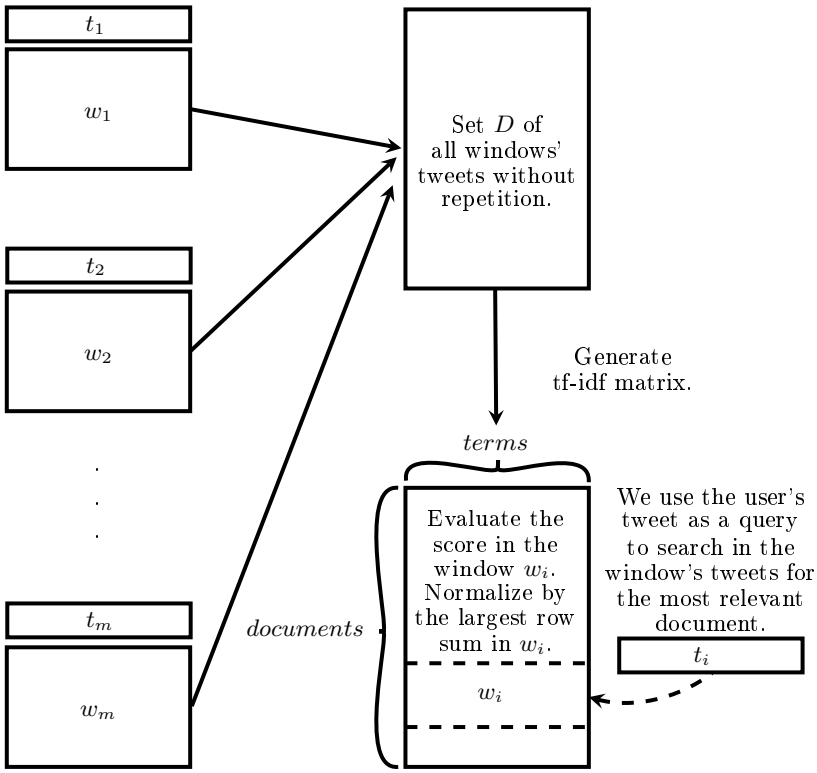



**Figure 2.2:** *Process to generate each tweet score. All the tweets in the windows are used to compose the corpus from which the tf-idf matrix for a given user is generated. Each user's tweets are then used as queries to search in their windows for the most relevant followee's tweet.*

### 2.3.2 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[1] 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 2.3. Finally, the tweet author's username is added as a feature since that is also visible in the feed.

### 2.3.3 Message Scoring

The text similarity metric used for this task was the *tf-idf* scoring. It is a proven technique [SB88, AT05] 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 [Rob04]. For a given $(document, term)$ matrix entry, the $tf$ function is the *term* occurrence count in the *document*, as in 2.1, and the *idf* function is given by Equation 2.2.

[1] Natural Language Toolkit.

$$tf(d, term) = \frac{|\{w|w \in d, w = term\}|}{|d|} \tag{2.1}$$

$$idf(D, term) = \log \frac{|D|}{|\{d|d \in D, term \in d\}|} \tag{2.2}$$

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 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 2.3.

$$tf\text{-}idf(document, term) = tf * (idf + 1) \tag{2.3}$$

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 2.4.

$$pairScore(t_i, ft_j) = \sum_{term \in (t_i \cap ft_j)} tf\text{-}idf(ft_j, term) \tag{2.4}$$

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 2.5.

$$normalization(w_i) = \max_{t \in w_i}(pairScore(t, t)) \tag{2.5}$$

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 2.6.

$$score(t_i) = \max_{ft_j \in w_i} \frac{pairScore(t_i, ft_j)}{normalization(w_i)} \tag{2.6}$$

The interpretation of the $score(t_i)$ is how likely $t_i$ is to be a response to a friend's tweet $ft = argmax_{ft' \in w_i}(pairScore(t_i, ft'))$ given the window $w_i$. It is important to notice that **this is not a pairwise tweet similarity**, but a **tweet-window similarity**, meaning that the similarity of one particular tweet in the window with $t_i$ might change depending on which other tweets are present in the window. This is a consequence of using tf-idf scoring and our normalization approach.

### 2.3.4 A Qualitative Understanding of the Score Method

In order to better understand our normalization technique, let's consider that we have a single window that contains two tweets, as in Table 2.4. In this case, if a user were to retweet tweet #1, he would get a final score of 0.333, since this is not the maximum possible tf-idf score in the window. On the other hand, if a user were to retweet tweet #2, his score would be 1. That is because the maximum tf-idf score in the window would be achieved only when retweeting tweet number #2. The reason why this makes sense is because if a tweet has fewer words, it is more likely that these

words are present in tweets that belong to the window just by chance. Therefore, it takes a longer sequence of matching words for a pair tweet/window to be high scored, while fewer matching words are a weaker signal. In other words, a high score would mean we have a good precision in identifying a retweet belonging to a window, but not necessarily a high recall. Showing that we can identify implicit reactions with a high precision method, even though it does not have an ideal recall, is enough for the proposed research questions in this work.

| Id | Tweet | Retweet tf-idf | Retweet score |
|---|---|---|---|
| 1 | `Nice Work!` | 0.301 | 0.333 |
| 2 | `This is a nice day to walk in the park.` | 0.903 | 1.0 |

**Table 2.4:** *Example of window. We do not consider stop words to calculate the tf-idf matrix.*

## 2.4 Results

### 2.4.1 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 2.5 provides an overview of the dataset and the generated windows, while Figure 2.3 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[1].

### 2.4.2 How prevalent are non-explicit responses?

This section addresses the first research question **Ch.2 RQ1** 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 2.6. This can also be seen in the scores' histogram for each of these sets in Figure 2.4. The

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

| | |
|---|---|
| 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 2.5:** *Overview of the data considering our window definition. 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.*

| | 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 2.6:** *Sample mean and standard deviation for the normalized similarity score for the Replies, Retweets, and Non-Tagged sets.*

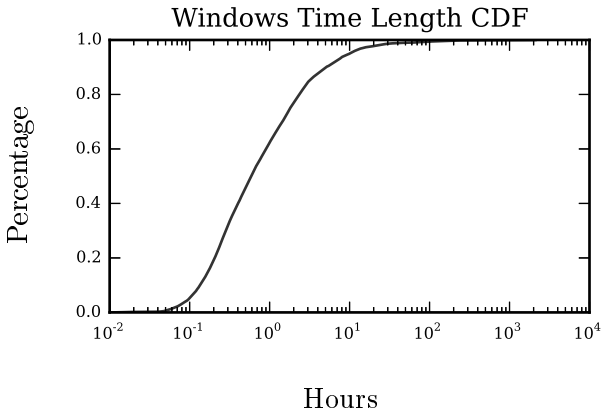



**Figure 2.3:** *Cumulative distribution function for the time length of the windows given in hours. Most windows' lengths are in the interval* $[10^{-1}, 10^{1}]$ *hours; about 60% of windows are 1 hour or less, meaning users receive on average over 100 tweets an hour.*

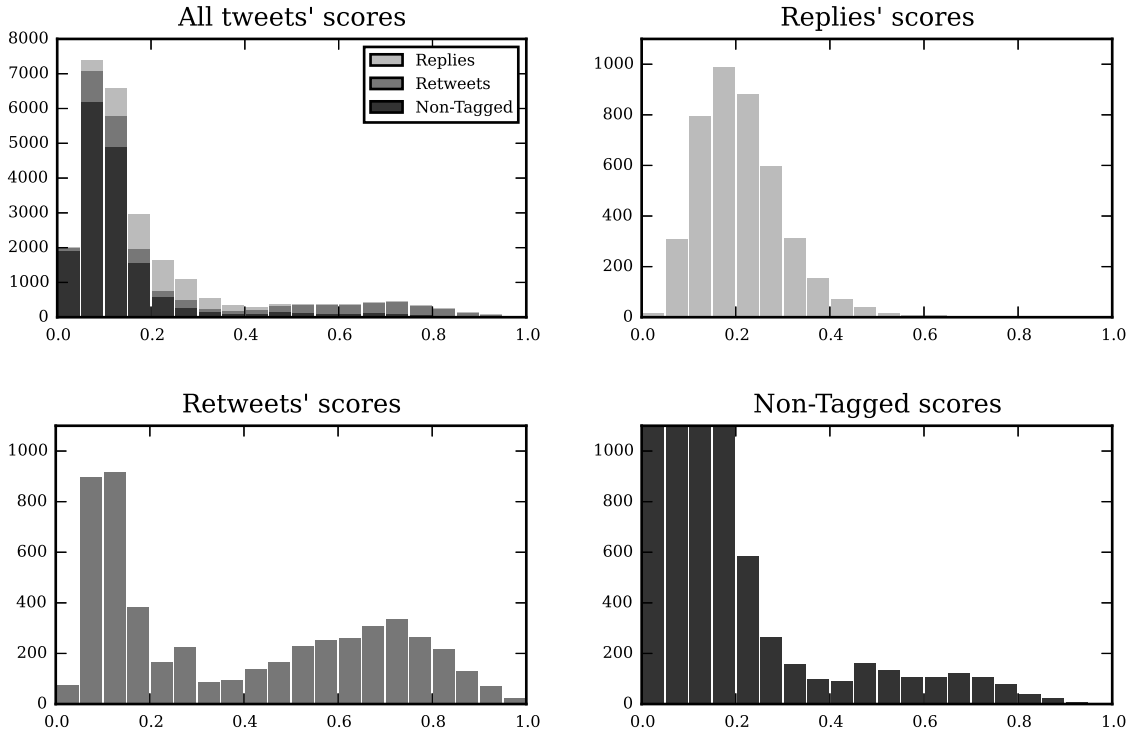



**Figure 2.4:** *Histograms for the normalized similarity scores. Note that the y-axis for the Non-Tagged subgraph was truncated at 1100 for better visualization of the tail of the distribution and matching other scales. Retweets have a higher average score than Replies, which in turn are higher than 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.*

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

**Table 2.7:** *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.*

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 2.4. Also, the Non-Tagged high scored tweets are not neglectable when compared with the number of high scored Tagged tweets, as seen in Table 2.7: 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 2.4. 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.

### 2.4.3 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. Tables 2.8, 2.9 and 2.10 show 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

| # | Score | Retweets |
|---|---|---|
| 1 | 1.0 | **User_8:** RT @User_21: 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. |
| | | **User_21:** 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 | **User_18:** RT @User_1: A girlfriend would be great, but I'm already in a pretty committed relationship with alcoholism and bad decisions. |
| | | **User_1:** A girlfriend would be great, but I'm already in a pretty committed relationship with alcoholism and bad decisions. |
| 3 | 0.768 | **User_19:** RT @User_20: All the weird horny stuff between Glenn and Maggie on Walking Dead makes me very uncomfortable. |
| | | **User_20:** All the weird horny stuff between Glenn and Maggie on Walking Dead makes me very uncomfortable. |
| 4 | 0.602 | **User_2:** RT @User_4: Brandywine Came Out With Win, #TeamBwine |
| | | **User_4:** Brandywine Came Out With Win, #TeamBwine |
| 5 | 0.522 | **User_6:** RT @User_22: I just love Toy Story, all of them |
| | | **User_22:** I just love Toy Story, all of them |
| 6 | 0.408 | **User_3:** RT @User_13: Welp now you know |
| | | **User_13:** Welp now you know |
| 7 | 0.303 | **User_25:** RT @User_16: Facebook is down? Oh no, how are cancer and child abuse going to stop without all those likes? :( |
| | | **User_11:** RT @User_9: Stop acting like we have 'rights' on Facebook http://t.co/geE9NjHH |
| 8 | 0.248 | **User_6:** RT @User_15: going to be slightly awkward when Jahmene scans the xfactor winner's single in Asda |
| | | **User_7:** RT @User_17: Jahmene is gonna be scanning James Arthers CD at ASDA now.. awks |
| 9 | 0.132 | **User_14:** RT @User_23: RG3 should give Michael Vick a class in scrambling. |
| | | **User_12:** RG3 running at 4G! |
| 10 | 0.070 | **User_10:** RT @User_24: It's Monday, User_24! http://t.co/asOF9yPA |
| | | **User_5:** Mondays are like Zubats. Nobody likes Zubats. |

**Table 2.8:** *Pairs of users' tweets (top in each row) and highest scoring messages in the windows (bottom in each row) for Retweets. Tweets were randomly selected across the range of scores in each set.*

| # | Score | Replies |
|---|---|---|
| 1 | 0.768 | **User_29:** RT @User_58: On deck - Next week's soup is White Bean and Smoked Turkey Chili! |
| | | **User_58:** On deck - Next week's soup is White Bean and Smoked Turkey Chili! |
| 2 | 0.693 | **User_53:** @User_45 @User_48 @User_32 @User_31 @User_47 @User_56 @User_49 Hate them. |
| | | **User_32:** @User_31 @User_47 @User_48 @User_56 @User_49 @User_45 @User_53 Hateful. Just hateful. |
| 3 | 0.573 | **User_30:** @User_27 @User_36 @User_33 @User_40 @User_51 @User_35 yeah! thanks emmit! muah! love, Hugs! for you! |
| | | **User_27:** @User_30 @User_36 @User_33 @User_40 @User_51 @User_35 happy birthday hector sending you love and hugs buddy. |
| 4 | 0.477 | **User_59:** @User_54 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 |
| | | **User_54:** RT @User_59: @User_54 lets do that thing were we get drunk and drive around the city while I'm playing my guitar ......again. |
| 5 | 0.386 | **User_50:** @User_57 yeah! Buzzin |
| | | **User_57:** @User_50 do you? :( |
| 6 | 0.283 | **User_52:** @User_52 @User_46 forgot the x hahaa |
| | | **User_46:** @User_52 aw hen, I feel for you x |
| 7 | 0.245 | **User_44:** @User_42 good morning! |
| | | **User_43:** Good morning everyone. |
| 8 | 0.168 | **User_39:** @User_37 HAPPY BIRTHDAY!:) xx |
| | | **User_34:** @User_39 i can be!! |
| 9 | 0.133 | **User_41:** @User_38 Niiiiice... |
| | | **User_38:** @User_41 lmao! I'm gonna put a slice of bacon with/in my drink and see what happens lol |
| 10 | 0.068 | **User_28:** @User_26 There it is. |
| | | **User_26:** RT @User_55: I am not a slut. I'm an erection enthusiast. |

**Table 2.9:** *Pairs of users' tweets (top in each row) and highest scoring messages in the windows (bottom in each row) for Replies. Tweets were randomly selected across the range of scores in each set.*

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.

| # | Score | Non-Tagged |
|---|---|---|
| 1 | 0.920 | **User_70:** RT @User_78: FLASH: #Egypt's Mursi has left presidential palace, two presidency sources say after protesters, police clash outside.<br>**User_63:** RT @User_78: FLASH: #Egypt's Mursi has left presidential palace, two presidency sources say after protesters, police clash outside. |
| 2 | 0.884 | **User_70:** It's time. RT @User_84: hey guys - this is barack. ready to answer your questions on fiscal cliff & #my2k. Let's get started. -bo<br>**User_68:** RT @User_84: hey guys - this is barack. ready to answer your questions on fiscal cliff & #my2k. Let's get started. -bo |
| 3 | 0.782 | **User_73:** A white woman... RT @User_81: A woman? RT @User_71: 'Fresh Prince' Star Alfonso Ribeiro Weds http://t.co/IfT3Zqlr<br>**User_81:** A woman? RT @User_71: 'Fresh Prince' Star Alfonso Ribeiro Weds http://t.co/iTrfZfeM |
| 4 | 0.697 | **User_85:** RT @User_83: hear @User_76 on the radio one minute ago!! it was funny :D x<br>**User_83:** hear @User_76 on the radio one minute ago!! it was funny :D x |
| 5 | 0.579 | **User_66:** RIP Mr Brubeck. Take five. "@User_62: Dave Brubeck, jazz icon, dead at 91. http://t.co/ae9UIRmP"<br>**User_80:** RT @User_62: Dave Brubeck, jazz icon, dead at 91. http://t.co/sOrBOFBR |
| 6 | 0.443 | **User_86:** @User_64 OKC traded James Harden<br>**User_60:** Why WOULD OKC TRADE JAMES HARDEN???? |
| 7 | 0.359 | **User_79:** (that should have had a link to the Tina and Amy host Golden Globes article. But I'm too lazy to fix it now)<br>**User_72:** Genius Move: Tina Fey and Amy Poehler to Host 2012 Golden Globe Awards! http://t.co/zdc2hS8F |
| 8 | 0.275 | **User_77:** @User_74 are you excited to come to Australia and meet all your amazing fans like me? ;) xx #asknialler 11<br>**User_61:** @User_75 EXCITED TO COME BACK TO AUSTRALIA, cause we miss you lots xox #asknialler |
| 9 | 0.242 | **User_79:** All of @User_69's #captainhottie pirate puns for #ouat are perfect. It's the reason we are twitter friends.<br>**User_69:** I apologize in advance for inappropriate pirate puns. #sorryImnotsorry #OUAT |
| 10 | 0.139 | **User_67:** Forever my lady lol<br>**User_65:** @User_82 just go get it!! Lol (the devil) aint I? Lol but I would lol |

**Table 2.10:** *Pairs of users' tweets (top in each row) and highest scoring messages in the windows (bottom in each row) for Non-Tagged tweets. Tweets were randomly selected across the range of scores in each set.*

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 "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.

### 2.4.4 Variations in User Responsiveness

The previous sections demonstrate that 11% or more of Non-Tagged tweets are likely to be responses that are not explicitly captured by the system. This section addresses the other main research question **Ch.2 RQ2** of how these losses are distributed among different users in the network.

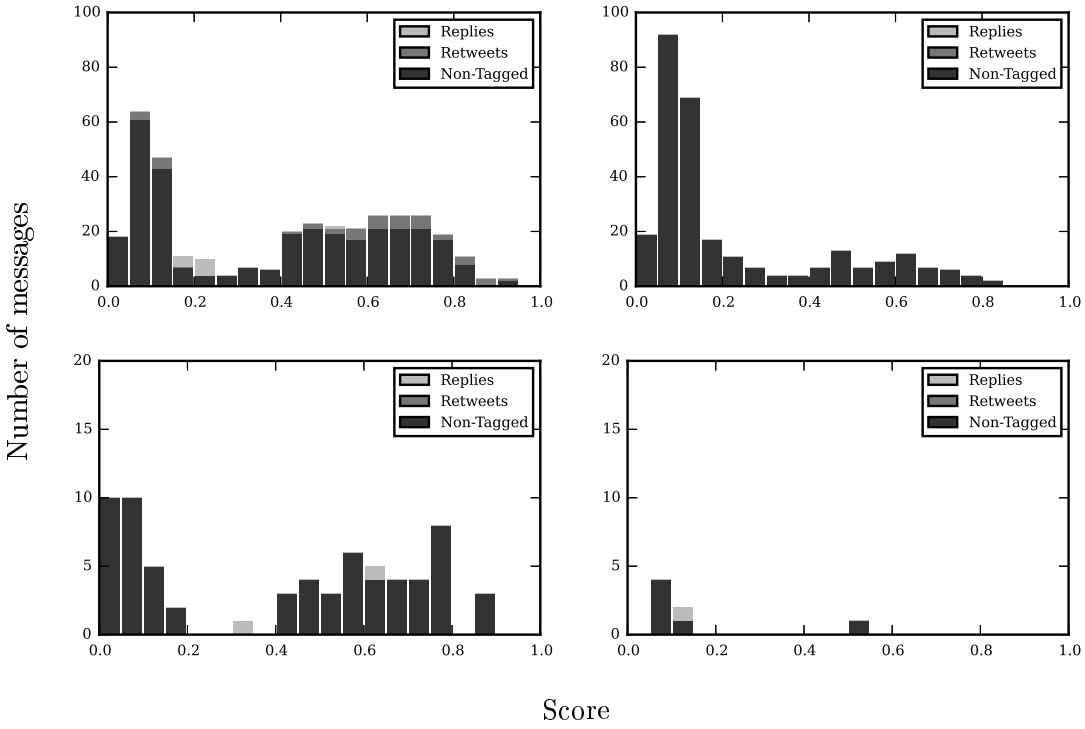



**Figure 2.5:** *Score histograms for sample users who present a significant amount of high scored Non-Tagged content relative to their total amount of messages, which indicates that most of their reactions are not being properly tagged by Twitter.*

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 2.5 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 do not know how to use the interface.

In order to better understand the behavior distribution among all users, Figure 2.6 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 this 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.

When considering the cumulative distribution of the users according to the percentage of high scored Non-Tagged messages $p_i^N$, shown in Figure 2.7, 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.

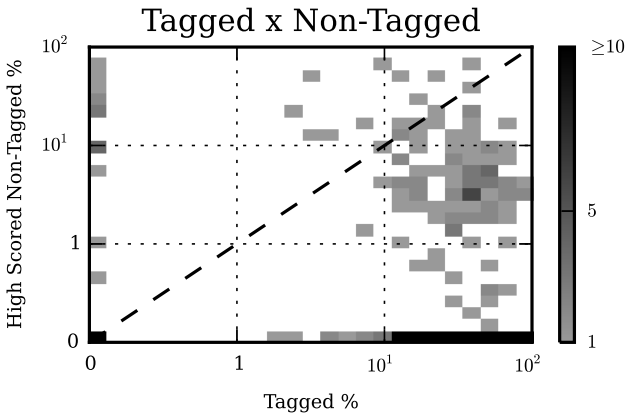



**Figure 2.6:** *2D histogram of the percentage of Tagged and high-scored Non-Tagged messages for all users. The scale is linear in the interval* $[0, 1]$ *and logarithmic on the interval* $(1, 100]$*; the dashed line represents an equal percentage of Tagged and Non-Tagged tweets. Many users are non-responsive (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).*

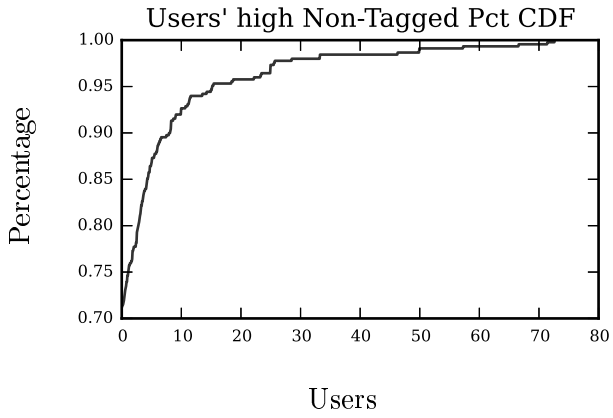



**Figure 2.7:** *Cumulative distribution of the users for the percentage of high scored Non-Tagged messages. 71% of the users have no high scored Non-Tagged messages, while 8% of the users had at least 10% of their messages high scored and Non-Tagged.*

### 2.4.5 Manual Retweeting

A significant number of high scored messages identified by our method were actual retweets that were not tagged by the system, meaning that users actually copied and pasted the original tweet in order to produce these particular retweets (often adding extra comments of their own). This is actually a common practice advised by online marketeers to increase reach and visibility of one's retweets [Sho, Joc, Fit]. Their main arguments is that the built-in retweets suffer a series of shortcomings: they lack the possibility of being edited by the retweeting user, followers can set their feed to filter out built-in retweets, trackings for the link interactions for a built-in retweet are attributed to the original post link only, lack of information about who retweeted, since it shows more information about the original poster. These issues might be reason enough to explain why we see more than 8% of users putting on the extra effort of manually coping and pasting, as well as editing the original message, in order to avoid these shortcomings of the built-in retweets.

### 2.4.6 Implicit Response Detection Through Provenance

Information provenance is a related field to the work presented here. Its main goal is to establish the origin of a given content. It can be useful to asses influence and trustworthiness, for instance. It is also interpreted as the reverse process of information diffusion [TDNV+15].

A relevant provenance model that is closely related to our work was published later by De Nies et al. [DNTD+15]. It uses a multi-level provenance method, and one of its steps to cluster messages in a tf-idf feature space. The clustering method chooses an arbitrary threshold that is closely related to the cluster size (higher thresholds lead to smaller clusters and vice-versa). For each cluster, the oldest message is said to be the root message of the cluster, and all the messages in the cluster share some provenance with it. This method was later used by Taxidou et al. [TFDN+16] to infer implicit interactions. Their work defined three possible types of interaction, which included subtypes: 1) user influence: a) with explicit credit, b) without credit, 2) external influence, 3) self-influence: a) delete and rewrite, b) promotion. From a total of 3909 messages, they found that explicit interactions were 2068 retweets, 198 quotes, and 93 replies, for a total of 2359 Tagged messages. Using the tf-idf similarity method and a threshold of 0.4 (for which they claim that messages above this similarity are likely to share some provenance), they detected another 192 messages that shared provenance but were not explicit Tagged. Using the same relative analysis as we did in our work, the implicit interactions represents about 8.1% of the total of explicit interactions.

When comparing our results, we should bear in mind that there are limitations to how these methods work and the datasets that were used. Their method takes a full dataset and cluster over all its messages, in contrast to our approach that is user-centered. Their dataset was crawled from the Twitter streaming API based on keywords related to the ISWC 2015 conference, therefore it is biased to this topic. In our dataset, we chose a set of users that followed Obama and crawled all their ego-network tweets, reducing possible biases since the bias would be in the choice of the user and not of the messages. Also, since it is not a user-centered approach, all analysis are comparable only at the message level, not allowing us to compare user's profiles and patterns of behavior. Finally, they considered quotes as an explicit reaction (which was included in the 8.1% previously mentioned), while our method only considered replies and retweets.

Regardless of these limitations, the fact that the underlying method uses tf-idf to evaluate similarity over tweets and identifies implicit reactions allow us to compare results. In our method, we identified 11% of the Non-Tagged tweets as possible reactions in comparison with the Tagged ones, opposed to 8.1% using their method. We can attribute this difference to the fact that our model takes into consideration a more specific baseline of comparison, since we calculate new tf-idf scores for each ego-network. Also, we used a different threshold for our score, which could yield these differences. Nevertheless, their results are in line with our first research question **Ch.2 RQ1**. It supports the claim that text similarity has the potential to detect implicit responses and presents estimates for theses numbers that are within the same order of magnitude of our results.

## 2.5 Contributions

This chapter presented a novel method for user's non-explicit reactions to followees' content detection in Twitter. It is based on text similarity scores between a user's tweets and those of their followees [BCJC15]. Our method generates higher scores on average for system tagged Replies and Retweets than Non-Tagged tweets, suggesting that our text similarity scores have the potential to identify users' reactions as stated in our first research question **Ch.2 RQ1**.

When using a conservative cutoff to detect non-tagged reactions, our results indicate that at least 11% of users' reactions are not tagged by the system. Considering our second research question **Ch.2 RQ2**, almost a quarter of the users in our dataset presented non-tagged reactions. Furthermore, among these users we have another quarter of them exhibiting more missed reactions than explicit system tagged ones. These users have a wide range of activity level, with dozens or hundreds of tweets in a 14-day window, meaning that they are not just naive, low-activity users who do not understand Twitter.

### 2.5.1 Limitations and future work

The pattern of behavior of each individual is taken into account in this work when we consider each users' ego-network for our model. However, we modeled reactions in a broad meaning: users could be reacting to each other, or reacting to a common external event or even just posting the same thing by coincidence. While all these effects are interesting and we would like to be able to distinguish users in these dimensions – users that have a high reaction rate to other users might be interesting as followers. Users that are prone to react to certain external events might be interesting as a target audience for these events. Users that just happen to post the same things might be clustered together for marketing campaigns. Our method as designed does not account for these differences. Once we detect a reaction, we could further refine it in these dimensions and have a more fine grained model of each user. This would be a natural development of this model.

Another important aspect that we did not consider here is that text is not everything. Nowadays more than ever, images and videos are commonly used in social networks posts and they carry a significant amount of information. Mapping and understanding how users react to this content could improve our method, and researchers such as Cavalin et al. [CFBP16] have been working in tools to better aggregate and understand Twitter posts that contain images.

# Chapter 3

# Averaging Gone Wrong: Using Time-Aware Analyses to Better Understand Behavior[1]

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 general patterns of behavior [ZKK14, KGM10], while others adopt a 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: datasets may 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 may be small [LKC08]. Aggregate time-based analyses are also a natural first way to address questions of community evolution.

However, we argue that many of these aggregated views are misleading. The conditions under which users join the community may vary greatly over time in ways that might impact their behavior [MCT15]. Among other things, popularity, purpose, features, interface, and algorithms can change: Wikipedia circa 2005 and circa 2015 are very different, as are Facebook of 2005 and 2015. Analyses—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 date of a user's first visible activity. We also look at differences within yearly cohorts, focusing on how behavior differs between shorter and longer-lived users within each cohort [BCSC16].

We find that these accountings for time reveal 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 versus users who leave Reddit early; none of these findings are obvious from aggregate views of user behavior.

[1] The contents of this chapter were published at the 2016 ACM 25th International Conference on World Wide Web (WWW'16)[BCJC15]. Some adaptations were made to include further research and adjust the format to this thesis.

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 (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 by lower-activity users leaving early.

We see the second part of this thesis 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.

## 3.1 Time matters

### 3.1.1 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, and social norms and structure. One aspect of time often considered is the tenure of a user in the community, as in studies around modeling users' 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], while their early experiences and activity shape future outcomes predictably [TL15, YC09, PHT09, MCT15].

However, much past work on online communities ignores the time at which a user joins the community and analyzes all users together. This might be a mistake: communities may grow denser or sparser with time [LKF05], develop new norms [KGM10], and 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 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 due to changes in the community, rather than any significant change in the outcome variable of interest or the user population.

### 3.1.2 Cohorts are analytically useful

A common method to control for such confounds 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. People in a cohort are assumed to be exposed to the same state of the world and thus are more comparable to each other than to 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 decision-making [AW93] or opinions on issues [FD88, Jen96]. Similarly, medical studies interpret effects of drugs using cohorts of people within the same age group or amount of exposure to correlated conditions [HECR96, DAE+10].

Recent work shows that cohorts' importance 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, 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].

### 3.1.3 What might cause these differences?

These differences in activity between cohorts may be due to a number of reasons. One plausible explanation is 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]. In this case, users who join earlier might be expected to be more active, committed users than those who join later.

Another possible explanation is that community norms may change over time. In many cases, it is a bottom-up process. Kooti et al. showed that social conventions can define the evolution of a community and the early adopters play a major role in designing these conventions, consciously or not [KGM10]. 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 between 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 users' experience. 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]. 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 come online, their influx may lead to changes in activity patterns and communities (as with the yearly entry of college freshmen, and eventually all of AOL, gaining access to Usenet). The gradual penetration of technology also has age-related effects: people who did not grow up in a technological environment differ in their social media and search usage compared to younger generations[CHdZ10, Bel05].

### 3.1.4 Is Reddit getting "worse" over time?

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]. Further, they suggest a general hypothesis that communities "get worse" over time because newer users are likely to be less committed and knowledgeable about the community.

To address this hypothesis, we analyze both aggregate and cohort-based measures of user quality that are often raised 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, a community that has been studied by many researchers [Gil13, Sto15, Ber11, TL15]. We begin with a brief overview of both Reddit and the dataset that we use in this work, focusing on aspects that directly impact our analyses[1].

[1] There is more to say about Reddit itself (see [Reda]).

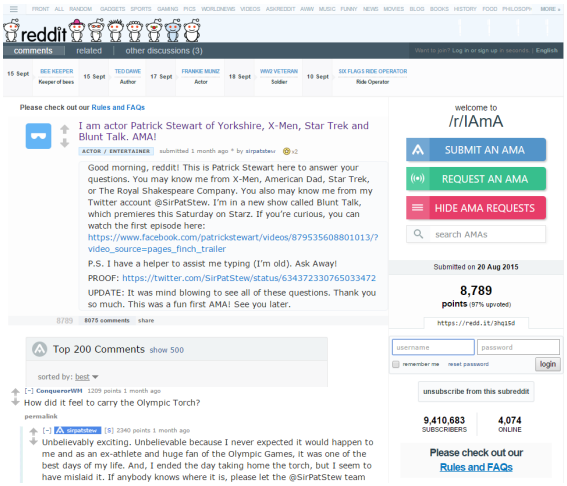



**Figure 3.1:** *Reddit interface when visualizing a submission. This is Patrick Stewart's "AmA" (ask me anything) in "IAmA" (I am a), a submission where he answers users' questions in the comments. We can see the most upvoted comment and Patrick's answer right below.*

## 3.2 Data: Reddit as a community

### 3.2.1 What is Reddit, briefly

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 June 21st, 2015[1]), 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", and "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 submission can be imagined as the root of a threaded comment tree, in which Redditors can comment on submissions or each other's comments. Redditors can also vote on both submissions and comments; these votes affect the order in which submissions and comments are displayed and also form the basis of "karma", a reputation system that tracks how often people upvote a given Redditor's comments and submissions. We can observe these elements in Figure 3.1.

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, a number of Reddit users believe that it is, in fact, getting worse over time[Pha, vor, God, ele, AND, Not]. Finally, Reddit data are publicly available through an API.

### 3.2.2 The dataset

Redditor *Stuck_In_The_Matrix* used Reddit's API to compile a dataset of almost every publicly available comment[Stu] 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 (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[2]; for our pur-

[1] [ra] provides more statistics about Reddit.
[2] A full description of the JSON objects is available at [Redb].

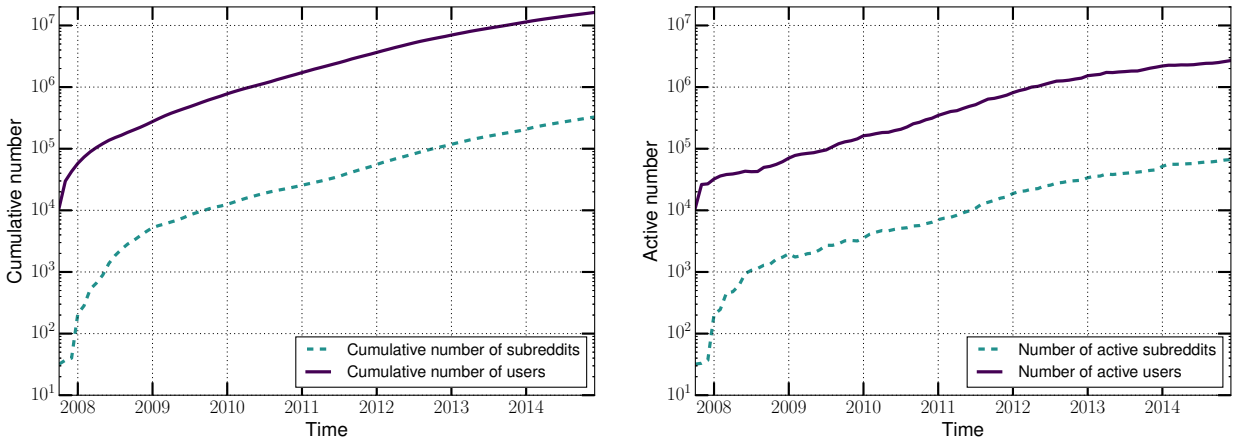



**Figure 3.2:** *Figure (a) shows the cumulative growth of Reddit for users and subreddits. Figure (b) shows the number of active users and subreddits in Reddit over time. An active user or subreddit is one that had at least one post (comment or submission) in the time bin we used—here, discretized by month.*

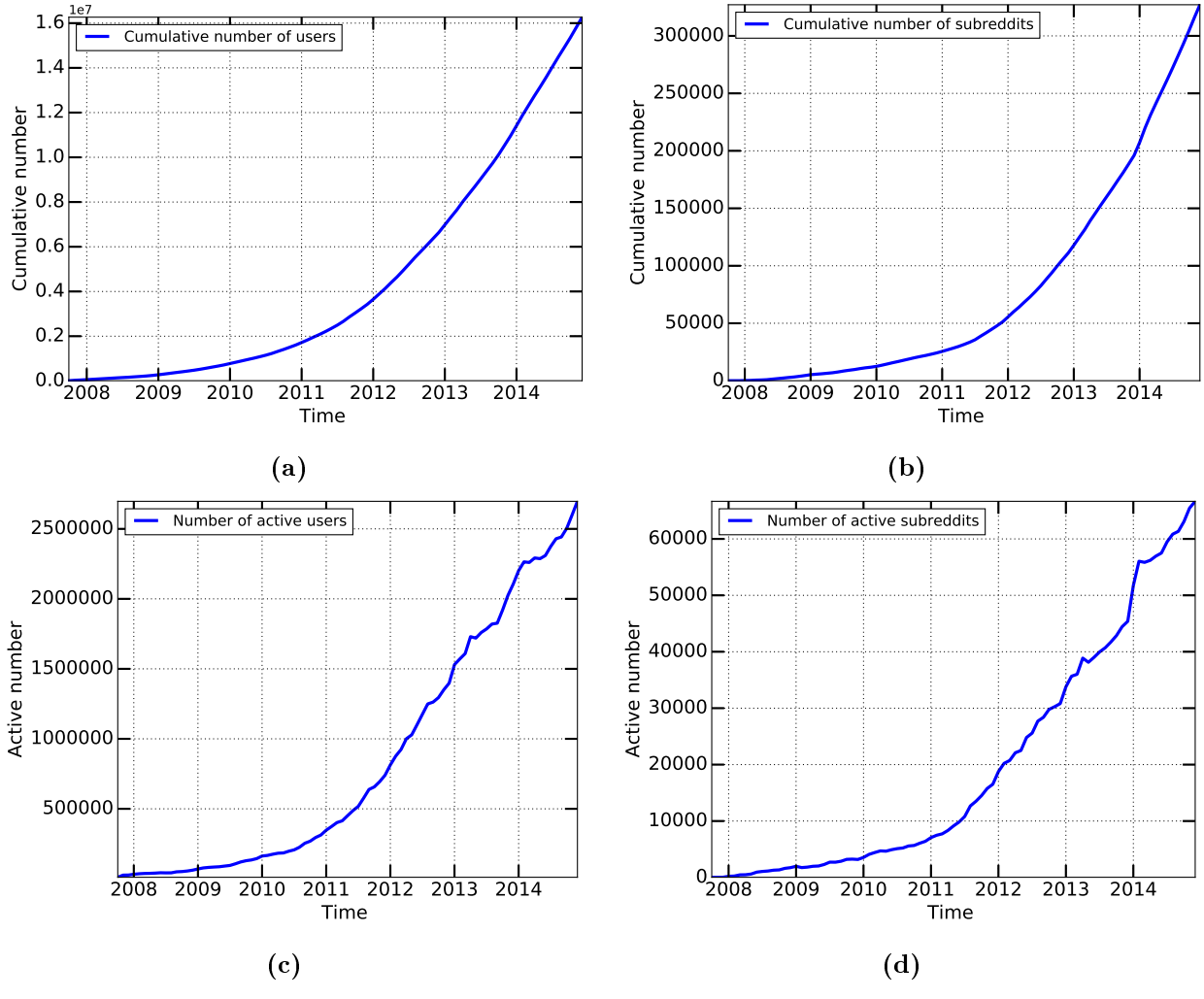



**Figure 3.3:** *These curves are the same as as those in Figure 3.2, but with the x axis in linear scale to highlight the exponential growth of reddit. Figure (a) is the cumulative number of users, (b) the cumulative number of subreddits, (c) the number of active users and (d) the number of active subreddits.*

poses, the main items of interest were the UTC creation date, the username, the subreddit, and for comments, the comment text.

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 time-based analyses. In some analyses, 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[1], but individual votes with timestamps and usernames are not available through the API, only the aggregate number of votes that posts receive.

### 3.2.3 Preprocessing the dataset

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

For the analysis in the work, we did light preprocessing to filter out posts by deleted users, posts with no creation time, and posts by authors with bot-like names[2].

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.

### 3.2.4 An overview of the dataset

Here we present an overview of the dataset that shows Reddit's overall growth. Figure 3.2a 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, the number of users and subreddits have grown exponentially. As of the end of 2014, about 16.2 million distinct users and 327 thousand subreddits made/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 2.7 million users and 66 thousand subreddits were active, both around a fifth of the cumulative numbers. Figure 3.2b shows the monthly number of active users and subreddits. In order to highlight the exponential growth of Reddit, Figure 3.3 shows the same curves as Figure 3.2, but in linear scale.

Our interest in this work 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.

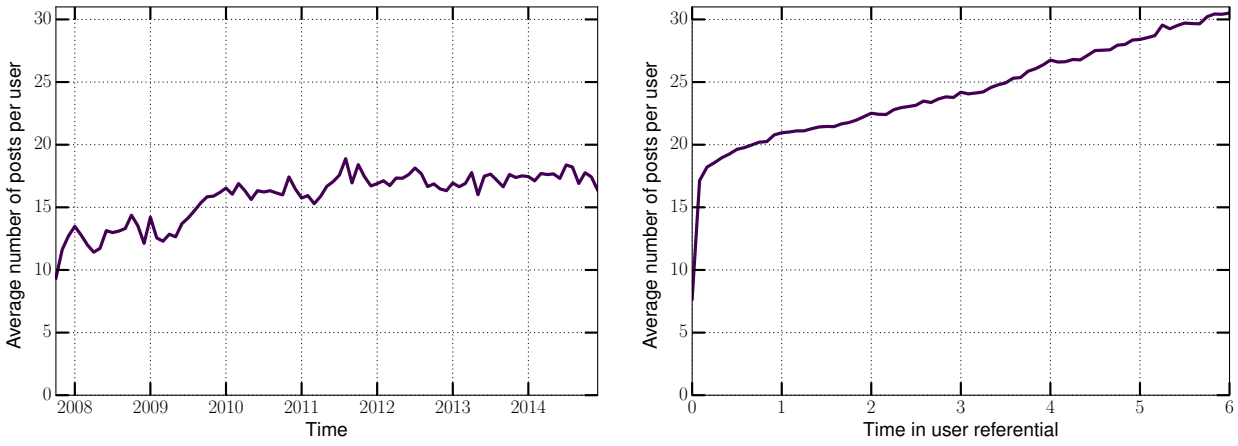
### 3.2.5 Identifying cohorts

We define the "**user's creation time**" as the time of the first post made by that user. Throughout this work, 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, meaning that not only do we have incomplete data for the 2007 year (which compromises this cohort), but there might also

[1] This would also give us more insight than usual into lurkers' behavior; we'll return to this in the discussion.
[2] Ending with "_bot" or "Bot"; or containing "transcriber" or "automoderator".




**(a)** **(b)**

**Figure 3.4:** *In Figure (a), monthly average posts per active user over clock time. In Figure (b), 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.*

be users and subreddits that show up in 2007 that were actually created in the previous years. 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 on the results; 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.

## 3.3 Average posts per user

One common way to represent user activity in online communities is quantity: the number of posts people make over time. Approaches that consider the total number of posts per user in a particular dataset [GLNGT04] and that analyze the variation of the number of posts per user over time [GTC+09] have been applied to online social networks. In this section, we use this measure to address our first research question (**Ch.3 RQ1**): how does the amount of users' activity change over time?

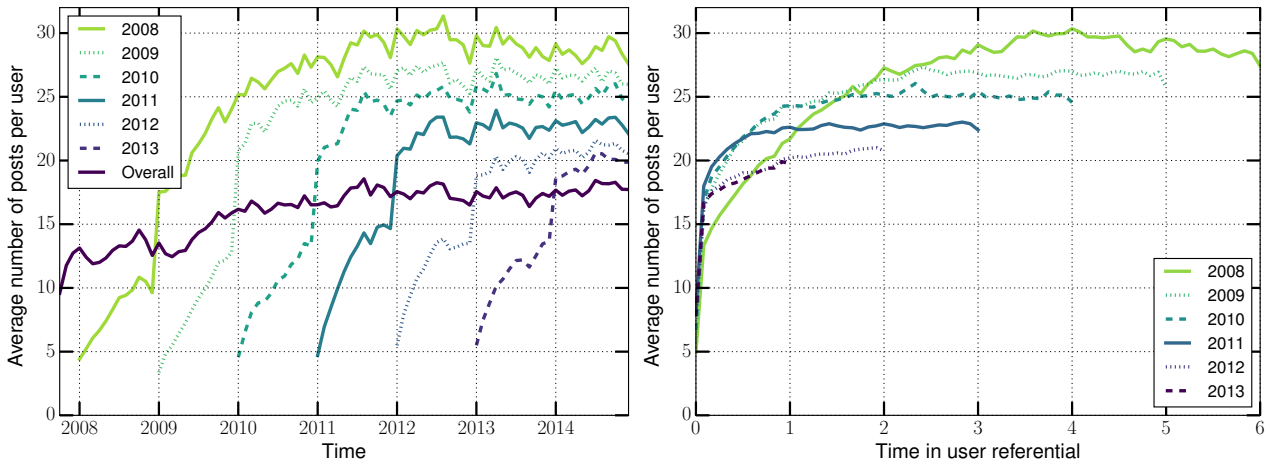
As we will see, both visualizing behavior relative to a user's creation time and using cohorts provide additional insight into posting activity in Reddit compared to a straightforward aggregate analysis based on calendar time.

### 3.3.1 Calendar versus user-relative time

Figure 3.4a shows that aggregate analysis, presenting 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]. Throw-away accounts are one example of very short-lived users in Reddit [Ber11], for example.

To address these characteristics, Figure 3.4b shows a view that emphasizes the trajectory over a user's lifespan rather than the community's. To do this, we scale the x-axis not by clock time, as




(a) (b)

**Figure 3.5:** *Figure (a) shows the average number of posts per active user over clock time and Figure (b) per active user 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) represents the overall average.*

in Figure 3.4a, 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. We call this the **time in the user referential**. One caution about interpreting graphs with time in the user referential 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.

The evidence at this point supports the tempting hypothesis that the longer a user survives, the more posts they make (**Ch.3 H1**). This hypothesis, however, is incorrect; we will present a more nuanced description of what is happening informed by cohort-based analyses.

### 3.3.2 New cohorts do not catch up

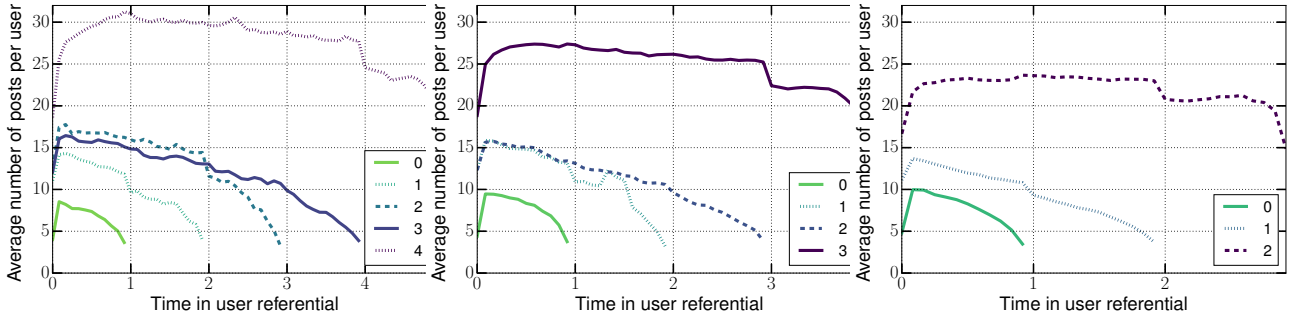
Figure 3.4b suggests that older users are more active than newer ones, raising the question of whether new users eventually follow in older users' footsteps (**Ch.3 RQ1a**).

Analyzing users' behavior by cohort is a reasonable way to address this question, and Figure 3.5a shows a first attempt at this analysis. We can already observe a significant cohort effect: users from later cohorts appear to level off at significantly lower posting averages than users from earlier ones. It suggests that newer users likely will never be as active as older ones on average. It also shows that surviving users are significantly more active than the overall average (the black line in the figure) would suggest.

However, Figure 3.5a also has an awkward anomaly: a rapid rise in the average number of posts during each cohort's first calendar year, especially in December. Combining cohort segmentation with user-referential analysis, as in Figure 3.5b, 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.

### 3.3.3 Does tenure predict activity, or vice versa?

These graphs still support our initial hypothesis **Ch.3 H1** and they do not explain the 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 (**Ch.3 H2**). 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 3.6:** *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 3.6 shows this analysis for the 2010, 2011 and 2012 cohorts[1]. 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 3.4b is not **Ch.3 H1**: longer-lived users don't post more as they age. Instead, users who post more—right from the beginning—live longer, supporting (**Ch.3 H2**).

Combining Figure 3.6'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.

## 3.4 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 put forward the following question (**Ch.3 RQ2**): how does comment length change in the community over time, both overall and by cohort?

### 3.4.1 Comment length drops over time

Figure 3.7a shows the overall comment length in Reddit over time (the darker line) and the overall length per cohort. Based on the downwards tendency of the overall comment length in Figure 3.7a, one might hypothesize that users' commitment to the network is decreasing over time (**Ch.3 H3**), or that there is some community-wide norm toward shorter commenting (**Ch.3 H4**).

However, this might not be the best way to interpret this information. Figure 3.7b shows the comment length per cohort in the user referential time. An important observation here is that younger users start from a lower baseline comment length than older ones. Considering the fact that Reddit has experienced exponential growth, the overall average for Figures 3.7a and 3.7b is heavily influenced by the ever-growing younger generations, who are more numerous than older survivors and who post shorter comments.

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

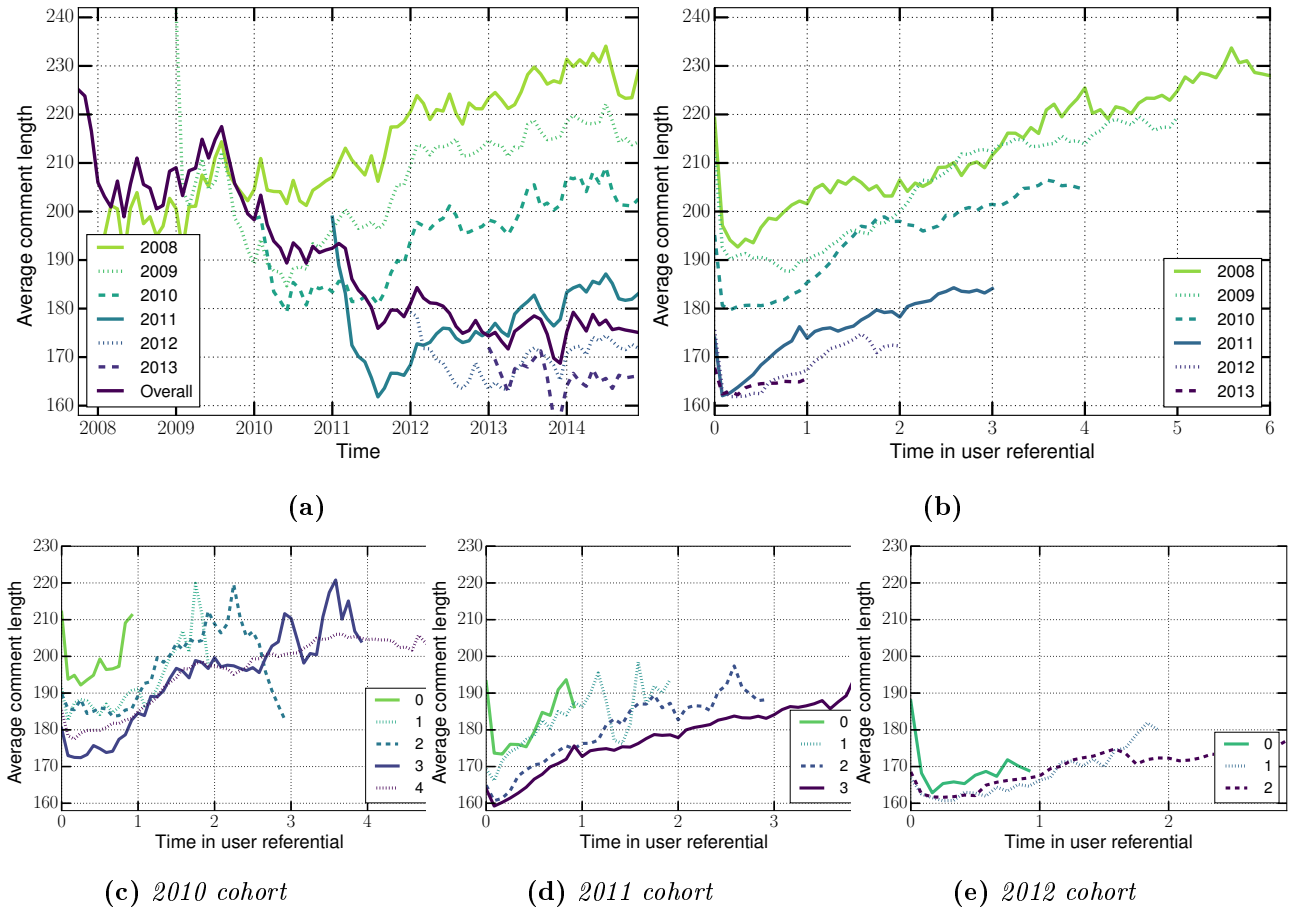



**Figure 3.7:** *Figure (a) shows the average comment length over clock time and Figure (b) from the user-referential time. Both figures show the cohorted trends. The overall average length per comment decreases over time, although for any individual cohort, it increases after a sharp initial drop. Figures (c), (d) and (e), similar to Figure 3.6, show 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.*

### 3.4.2 Simpson's Paradox: the length also rises

Let us go back to Figure 3.7a, 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, Figure 3.7b shows that average comment length increases over time for every cohort. While later cohorts start at smaller comment length, after an initial drop, on average all cohorts write 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.

Table 3.1 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.

| | 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 3.1:** *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 start generating data in their cohort year, therefore the upper diagonal is blank. On the right column we see the overall average for all users.*

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 3.7a.

Without the decision to condition on cohorts, one would have gathered an entirely wrong conclusion. People are not writing less as they survive, contra (**Ch.3 H3**). Rather, those who tend to write less are joining the community in much larger numbers. Why later users write less is an open question we speculate about later in the discussion and future work section.

### 3.4.3 New users burn brighter

As with the number of posts per user, we cannot say if the increase in the curves seen in 3.7b is due to lower-effort users dying first or because users are writing more as they live longer. The sub-cohort analysis in 3.7c allows us to make two observations toward this question. First, *comment length does increase inside of each cohort*, no matter how long the user survives. Second, as a general trend, *users 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.

## 3.5 Kinds of contributions

In addition to questions of effort, the online community literature also often asks what sorts of activities users engage in, for instance, to categorize users into roles they play in the community [WCK+11]. As with comment length, we propose the following research question (**Ch.3 RQ3**): how do users' activities change in the community over time, both overall and by cohort?

### 3.5.1 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. Submissions can be considered new content that an author generates, while comments can be considered as contributions toward existing content from another author.

Since the total number of comments always surpasses the number of submissions, we compute a user's ratio of comments per submission as a rough measure of the kinds of contributions they make. Figure 3.8a shows the overall and cohorted evolution of comments per submission from 2008 to 2013. Users who most prefer commenting to submitting come from 2009 to 2011, while over time the average ratio of comments to submissions increases both overall and per-cohort for active users.

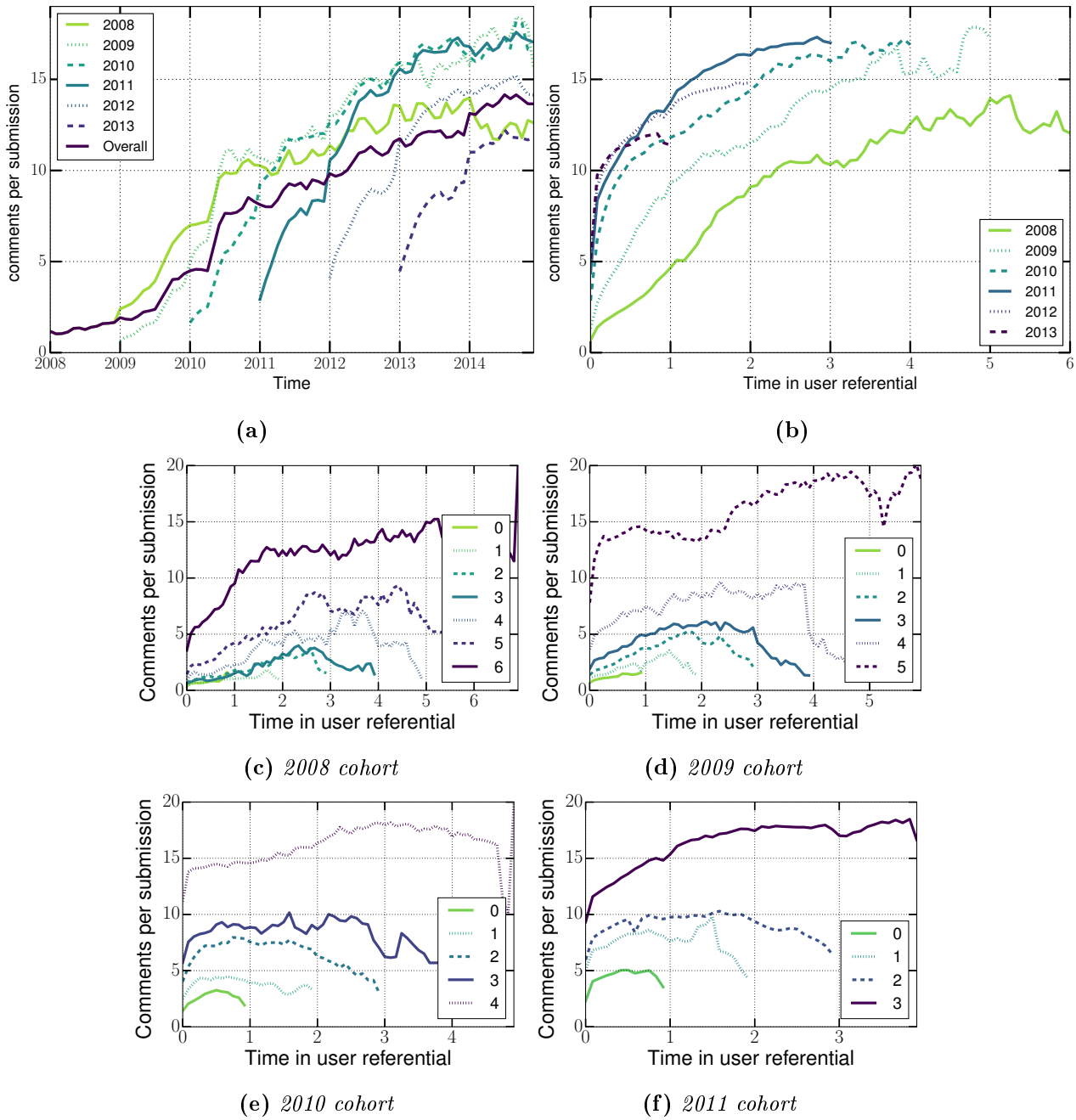



**Figure 3.8:** *Figure (a) shows the average comment per submission ratio over clock time for the cohorts and the overall average. Figure (b) shows the average comment per submission from the user-referential time for the cohorts. Figures (c), (d), (e) and (f), similarly to Figure 3.6, shows the 2008, 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).*

Again, we analyze our data from the user-time referential, as seen in Figure 3.8b. It shows a clear pattern for users in earlier cohorts to have a lower comment per submission ratio than users in later cohorts, 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).

### 3.5.2 Comment early, comment often

Figures 3.8c-f shows the cohorts from 2008 to 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 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 useful early predictor of user survival.

Second, and unlike the case for average post length, surviving users' behavior changes over time. For post length, Figure 3.6 shows that even the most active users come in at a certain activity level and stay there, perhaps even slowly declining over time. Here, Figures 3.8c-f show that 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 in the earlier cohorts of 2008 and 2009, with ratios more than doubling over their first year, much more than for later cohorts.

## 3.6 Discussion and Future Work

Our work highlight how common measures of activity and effort adopted by the research community can be misleading if it does not take into consideration the temporal evolution of the network. Other measures could be used, for instance, reddit has upvotes and downvotes, as well as a feature to give "gold". While we investigated these measures, our approach focused in examples that clearly showed counter intuitive results (particularly the ones related to the Simpson's paradox). Despite that other metrics can be studied all by themselves, the metrics we chose were the correct ones to show the desired effects: how changes in users' behavior over time can be hidden or misinterpreted when taken as an aggregate. They allowed us to answer our research questions, which we summarize below.

- **Ch.3 RQ1: How does the amount of users' activity change over time?**
  - Low activity users die sooner than high activity users. At the same time, the activity level is lower for newer cohorts than for older ones. This 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.
  - **Ch.3 RQ1a: Does new users eventually reach the higher level of activity of older users?**
    - Users from later cohorts appear to level off at significantly lower posting averages than users from earlier ones, suggesting that newer users likely will never be as active as older ones on average.
- **Ch.3 RQ2: How does comment length change in the community over time, both overall and by cohort?**
  - Comment does increase inside of each cohort, no matter how long the user survives. Users that make longer comments inside of each cohort die faster. However, the community-wide comment length decreases over time. This is due to the exponential influx of new users that start from a lower baseline of comment length in reddit, being an example of the Simpson's paradox.
- **Ch.3 RQ3: How do users' activities change in the community over time, both overall and by cohort?**

– Average comments per submission seem to be a good predictor of user survival in the community, as long-lived users start from a higher baseline in comparison to short-lived ones. Also, in every cohort, users increase their comment-to-submission ratio as they survive. As the total number of posts stays steady, this suggests that users substitute making their own submissions for commenting.

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.

### 3.6.1 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; several of our results support this explanation.

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.

### 3.6.2 Why are comments getting shorter?

We also observed that overall, comment lengths are getting shorter over time. One hypothesis is that users' behavior is being shaped by an "initial value problem"—that as users join the network, they tend to produce content according to the norms of what they see [KGM10, DNMWJP13]. Figure 3.7a presents some support for 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[rhi] 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 around modeling strategic posting and attention distribution behavior in Reddit.

### 3.6.3 Why do comments per submission increase?

We also saw that comments per submission increase over time for surviving users, especially 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 to comment on; those comments in turn provide additional opportunities for commenting. In this

reading, the value and ease of commenting has increased over time, making it a more common behavior.

This question of ease and value might be more general, and tie to our observations about self-selection and karma accumulation. Most users in social networks are known to be lurkers: seeking information and observing, rather than contributing content [RRS04, NP00]. Consumption in Reddit is valuable and easy, and 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 comments is relatively easy or relatively valuable are likely to be frequent submitters or "power users" [PHT09, KCP+07]. We suspect such users are more likely to be ones who found Reddit earlier, when it was relatively small, and stuck with it.

### 3.6.4 Limitations and Future Work

In this work we focused our attention on visible behavior attributable to specific users, which in this dataset meant submissions and comments. As with many analyses 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 use them as activity measures for specific users. The existing voting data might be much more useful, however, in addressing questions that involve predicting a given user's future behavior based on how other users respond to a user's early contributions [JK06, SWL12].

Focusing on visible activity can lead to blind spots in other places, as well. In particular, our emphasis on active users led us to ignore questions of survival, leaving, and rejoining. This was a reasonable view of the community based on the questions we were asking, 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; more generally, we see careful thinking about what it means to "survive" in a community as an interesting problem in its own right. Many analyses assume that a gap of some time period implies that a user has left, or that users "die" on their last visible day of activity. However, long gaps are common in real behavior. People temporarily quit social media all the time [BA13], and in Wikipedia, the practice of leaving temporarily is so common it has a name: "wikibreak". Rather than an annoying right censorship statistical problem, this question of what it means when contributors to a community start and stop 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 and leverage "bursty" ones.

Further, One of our assumptions was that one account is associated with one user. This might not be the case, as more than one user can share the same account [Lam14] or one user can have multiple accounts [Ber11]. Multiple accounts can have many functions, including making points someone doesn't want connected with their main identity, trolling or harming other users or the community, or simulating users who agree with a main identity ("sock puppets"). While we think this is not the main driver of our results, this should be checked in future work—and sockpuppet detection and account deanonymization is an interesting question in its own right.

Finally, focusing on visible activity can also lead to blind spots around deleted content or communities. At least in Reddit, activity from users is marked with a username of "[deleted]", which we discovered after realizing that one author had millions of comments(!), and that allowed us to consciously choose to exclude that data. However, in some contexts, such as Wikipedia articles that are deleted, that activity is invisible as edit behavior on those articles does not show up in many data dumps. Such invisible activity might be important in understanding either individual users or the community.

# Chapter 4

# Conclusions and Final Remarks[1]

In this thesis, we focus on users' reactions and evolution on social networks. In the second chapter, our efforts are directed at the problem of missed reactions. We propose a text similarity method to identify missed reactions and validate it on Twitter data. We estimate the potential missed behavior from users that are not captured by current Twitter mechanisms and we identified types of users that are significantly underrepresented based solely on these metrics. In the third chapter, we analyze Reddit users' evolution from a cohort perspective built on top of the time that they joined the network. We found wide differences in behavior depending on the year that they joined the network as well as how not account for temporal effects can be misleading. This thesis gives us better understanding of how we might ignore or misrepresent users' behavior on social networks and provide methods and analysis tools that help us to avoid such problems in the future. Here summarize our findings and some of their implications.

## 4.1 Reaction Detection

The first part of this thesis 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 [BCJC15]. 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 (**Ch.2 RQ1**). 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 (**Ch.2 RQ2**).

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. Further research could address 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

[1] The contents of this chapter are as seen in the works published during the course of this Ph.D.[BCJC15, BCSC16]. Some adaptations were made to include further research and adjust the format to this thesis.

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. Also, some practical applications for this kind of detection would include the identification of bots on social networks, since these tend to show a distinct pattern of behavior, as their reactions are likely different from a real human being. The matter of fake news could also be addressed if we could pinpoint some distinct patterns of reactions by the users.

## 4.2 Users' Behavior Evolution

The second part of this thesis 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 [BCSC16]. Although simple, this approach provides a number of insights that would be missed by straightforward aggregate analysis methods. We also analyze the evolution of users 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.

While analyzing how the amount of posting changes over time (**Ch.3 RQ1**), 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 (**Ch.3 RQ1a**). Controlling for survival provided evidence for hypothesis (**Ch.3 H2**), that users have a stable level of posting activity over time (with slightly decreasing patterns). Further, the percentage of surviving but low-activity users is increasing in the younger cohorts

When looking at changes in comment length over time (**Ch.3 RQ2**) as a proxy for users' effort, we found that while the overall average in Reddit seems to decrease, users actually write longer comments as they survive, no matter when they join. However, later cohorts of users that joined the network are writing smaller comments; their greater number leads to an instance of Simpson's paradox, where the overall average decreases while the series for each individual cohort increases.

Finally, we analyzed whether users change their commenting versus submission behavior over time (**Ch.3 RQ3**). We found that users with a higher initial comment to submission ratio survive longer on average, and that this ratio increases for surviving users, particularly for earlier cohorts. This isn't because activity rises overall, as posting activity remains stable; instead, it suggests that longer-term users substitute commenting for submissions.

An important remark of this thesis is how different demographics of users joining and leaving a network play a significant role in shaping the average user behavior. Failing to account for these might limit our interpretation of the data (**Ch.3 H1**, **Ch.3 H3** or **Ch.3 H4**) and lead to wrong conclusions.

Both our 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 works in this area.

# 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. 28

[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. 24

[AND] ANDYVDL. Reddit literally got worse under spez. https://www.reddit.com/r/PaoYongYang/comments/3gz6cb/reddit_literally_got_worse_under_spez/. 26

[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. 6, 7, 23

[AT05] Gediminas Adomavicius e Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005. 11

[AW93] Orazio P Attanasio e Guglielmo Weber. Consumption Growth, the Interest Rate and Aggregation. *Review of Economic Studies*, 60(3):631–49, 1993. 24

[BA13] Eric Baumer e Phil Adams. Limiting, leaving, and (re) lapsing: an exploration of facebook non-use practices and experiences. *Chi 2013*, páginas 3257–3266, 2013. 37

[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. 25

[BCJC15] Samuel Barbosa, Roberto M Cesar-Jr e Dan Cosley. Using text similarity to detect social interactions not captured by formal reply mechanisms. Em *e-Science (e-Science), 2015 IEEE 11th International Conference on*, páginas 36–46. IEEE, 2015. 4, 5, 20, 23, 39

[BCSC16] Samuel Barbosa, Dan Cosley, Amit Sharma e Roberto M. Cesar, Jr. Averaging gone wrong: Using time-aware analyses to better understand behavior. Em *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, páginas 829–841, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. 4, 23, 39, 40

[Bel05] S. Beldona. Cohort Analysis of Online Travel Information Search Behavior: 1995-2000. *Journal of Travel Research*, 44(November):135–142, 2005. 24, 25

[Ber11] Kelly Bergstrom. "don't feed the troll": Shutting down debate about community expectations on reddit.com. *First Monday*, 16(8), 2011. 25, 29, 37

[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. 6

[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. 24

[BKL09] Steven Bird, Ewan Klein e Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edição, 2009. 11

[BM11] D Boyd e Alice Marwick. Social privacy in networked publics: Teens' attitudes, practices, and strategies. Em *Symposium on the Dynamics of the Internet and Society*, páginas 1–29, 2011. 3

[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. 8

[BR09] Fabrício Benevenuto e T Rodrigues. Characterizing user behavior in online social networks. *Proceedings of the 9th ...*, páginas 49–62, 2009. 3

[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. 5

[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. 5

[CAKL10] Boreum Choi, Kira Alexander, Robert E Kraut e John M Levine. Socialization Tactics in Wikipedia and Their Effects. *Cscw*, páginas 107–116, 2010. 25

[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. xvii, 3, 5, 6, 7, 10, 13

[CFBP16] Paulo Cavalin, Flavio Figueiredo, Maira de Bayser e Claudio Pinhanez. Organizing images from social media to monitor real-world events. Em Daniel G. Aliaga, Larry S. Davis, Ricardo C. Farias, Leandro A. F. Fernandes, Stuart J. Gibson, Gilson A. Giraldi, João Paulo Gois, Anderson Maciel, David Menotti, Paulo A. V. Miranda, Soraia Musse, Laercio Namikawa, Mauricio Pamplona, João Paulo Papa, Jefersson dos Santos, William Robson Schwartz e Carlos E. Thomaz, editors, *Proceedings...*, Porto Alegre, 2016. Conference on Graphics, Patterns and Images, 29. (SIBGRAPI), Sociedade Brasileira de Computação. 21

[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. 23, 25

[CHK10] Dan Cosley, Daniel Huttenlocher e Jon Kleinberg. Sequential influence models in social networks. *... on Weblogs and Social ...*, páginas 26–33, 2010. 23

[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. 24

[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. 23, 24, 25, 36

[DNTD+15] Tom De Nies, Io Taxidou, Anastasia Dimou, Ruben Verborgh, Peter M Fischer, Erik Mannens e Rik Van de Walle. Towards multi-level provenance reconstruction of information diffusion on social media. Em *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, páginas 1823–1826. ACM, 2015. 6, 20

[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. 5

[ele] eleftheriaHthanatos. Is it just me or is reddit getting worse? https://www.reddit.com/r/Anarchism/comments/2qmhgf/is_it_just_me_or_is_reddit_getting_worse/. 26

[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. 24

[fho] fhoffa. 1.7 billion reddit comments loaded on BigQuery. https://www.Reddit.com/r/bigquery/comments/3cej2b/17_billion_Reddit_comments_loaded_on_bigquery/. 28

[Fit] Laura Fitton. How to Retweet the Right Way (With a Comment) on Twitter. https://blog.hubspot.com/blog/tabid/6307/bid/27675/How-to-Retweet-the-Right-Way-in-4-Easy-Steps.aspx. 19

[GAdS+13] Maira A de C Gatti, Ana Paula Appel, Cicero Nogueira dos Santos, Claudio Santos Pinhanez, Paulo Rodrigo Cavalin, Samuel Barbosa Neto, Maira A. De C Gatti, Ana Paula Appel, Cicero Nogueira Dos Santos, Claudio Santos Pinhanez, Paulo Rodrigo Cavalin e Samuel Barbosa Neto. A simulation-based approach to analyze the information diffusion in microblogging online social network. Em *Proceedings of the 2013 Winter Simulation Conference (WSC'13)*, páginas 1685–1696, 2013. 9

[GAP+13] Maira Gatti, Ana Paula Appel, Claudio Pinhanez, Cicero Dos Santos, Daniel Gribel, Paulo Cavalin e Samuel Barbosa Neto. Large-scale multi-agent-based modeling and simulation of microblogging-based online social network. Em *14th International Workshop on Multi-Agent-Based Simulation (MABS'13)*, 2013. 9

[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. 8

[Gil13] Eric Gilbert. Widespread Underprovision on Reddit. *Proceedings of the 2013 conference on Computer-supported Cooperative Work*, páginas 803–808, 2013. 25

[GK09] Eric Gilbert e Karrie Karahalios. Predicting tie strength with social media. *ACM Conference on Human Factors in Computing Systems*, páginas 211–220, 2009. 3, 5, 8

[Gle05] Norval D Glenn. *Cohort analysis*, volume 5. Sage, 2005. 24

[GLNGT04] D. Gruhl, David Liben-Nowell, R. Guha e A. Tomkins. Information diffusion through blogspace. *ACM SIGKDD Explorations Newsletter*, 6(2):43–52, 2004. 25, 29

[God] GodOfAtheism. [Meta][Hypothetical] How would you make reddit worse? [Brackets]. https://www.reddit.com/r/circlebroke/comments/19y889/metahypothetical_how_would_you_make_reddit_worse/. 26

[Goo] Google. BigQuery. https://cloud.google.com/bigquery/. 28

[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. 24

[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. 25, 29

[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. 6, 7

[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. 24

[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. 8

[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. 25, 28

[Ing14] Mathew Ingram. Digg Redesign Met With a Thumbs Down, 2014. 25

[Jen96] M. Kent Jennings. Political Knowledge Over Time and Across Generations. *Public Opinion Quarterly*, 60:228–252, 1996. 24

[JK06] Elisabeth Joyce e Robert E. Kraut. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11:723–747, 2006. 5, 37

[Joc] Adrian Jock. 5 Reasons Why Not to Retweet Using Twitter's Native Button. http://www.adrianjock.com/native-retweets-bad-marketing/. 19

[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. 25, 28

[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. 37

[KGM10] Farshad Kooti, Krishna P Gummadi e Winter a Mason. The Emergence of Conventions in Online Social Networks. *Artificial Intelligence*, páginas 194–201, 2010. 23, 24, 25, 36

[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. 25

[Lam14] Airi Lampinen. Account sharing in the context of networked hospitality exchange. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, páginas 499–504, 2014. 37

[Lar14] F Lardinois. Digg redesign tanks: Traffic down 26%(updated with new reddit stats), 2014. 25

[LH08] Xinxin Li e Lorin M. Hitt. Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474, 2008. 25, 36

[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. 6, 7

[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. 23

[LKF05] Jurij Leskovec, Jon Kleinberg e Christos Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. *In KDD*, páginas 177–187, 2005. 24

[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. 6, 7

[LW84] Mark R Levy e Sven Windahl. Audience activity and gratifications a conceptual clarification and exploration. *Communication research*, 11(1):51–78, 1984. 25

[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. 23, 24

[MF12] William M Mason e Stephen Fienberg. *Cohort analysis in social research: Beyond the identification problem*. Springer Science & Business Media, 2012. 24

[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. 24

[NGC+13] Samuel Martins Barbosa Neto, Maira Athanazio De Cerqueira Gatti, Paulo Rodrigo Cavalin, Claudio Santos Pinhanez, Cicero Dos Santos, Ana Paula Appel, Samuel Martins Barbosa Neto, Maira Athanazio De Cerqueira Gatti, Paulo Rodrigo Cavalin, Claudio Santos Pinhanez, Cicero Nogueira Dos Santos e Ana Paula Appel. Reaction times for user behavior models in microblogging online social networks. Em *Proceedings of the 2103 workshop on Data-driven user behavioral modelling and mining from social media (DUBMOD'13)*, páginas 17–20, New York, New York, USA, 2013. ACM Press. 4, 5, 9

[Not] NotaMethAddict. Reddit isn't getting worse at aggregating the news, everyone else has just gotten better. https://www.reddit.com/r/TheoryOfReddit/comments/3o9wk1/reddit_isnt_getting_worse_at_aggregating_the_news/. 26

[NP00] Blair Nonnecke e Jenny Preece. Lurker demographics: Counting the silent. *Proceedings of the SIGCHI conference on ...*, 2(1):1–8, 2000. 37

[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. 37

[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. 23

[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. 23

[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. 8

[Pha] Phallic. Posts about "how reddit is getting worse" are getting worse. https://www.reddit.com/r/reddit.com/comments/9y3yi/posts_about_how_reddit_is_getting_worse_are/. 26

[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. 23, 24, 25, 30, 37

[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. 6, 7

[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. 24

[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. 25

[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. 12

[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. 6, 7

[ra] reddit admins. Blog.reddit. http://www.redditblog.com/2015/06/happy-10th-birthday-to-us-celebrating.html. 26

[Reda] Reddit. About reddit. https://www.reddit.com/about/. 25

[Redb] Reddit. JSON. https://github.com/Reddit/Reddit/wiki/JSON. 26

[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. 37

[rhi] rhiever. Retracing the evolution of Reddit through post data. https://www.reddit.com/r/TheoryOfReddit/comments/1a7aoj/retracing_the_evolution_of_reddit_through_post/. 36

[Rob04] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004. 11

[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. 37

[Ryd65] Norman B. Ryder. The cohort as a concept in the study of social change. *Americal Sociological Review*, 30(6):843–861, 1965. 25

[SB88] Gerard Salton e Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988. 11

[SGC13] A Sharma, M Gemici e Dan Cosley. Friends, Strangers, and the Value of Ego Networks for Recommendation. *ICWSM*, 2013. 8

[Sho] Shoq. Why You Shouldn't Use Twitter's Built-In Retweet Feature Too Often. http://shoqvalue.com/dont-use-twitter-retweets. 19

[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. 5, 6, 7

[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. 25, 28

[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. 8

[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. 32

[Sto15] Greg Stoddard. Popularity and Quality in Social News Aggregators: A Study of Reddit and Hacker News. Em *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, páginas 815–818, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee. 25

[Stu] Stuck_In_the_Matrix. I have every publicly available Reddit comment for research. ~ 1.7 billion comments @ 250 GB compressed. Any interest in this? https://www.Reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_Reddit_comment. 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. 37

[TDNV+15] Io Taxidou, Tom De Nies, Ruben Verborgh, Peter M Fischer, Erik Mannens e Rik Van de Walle. Modeling information diffusion in social media as provenance with w3c prov. Em *Proceedings of the 24th International Conference on World Wide Web*, páginas 819–824. ACM, 2015. 20

[TFDN+16] Io Taxidou, Peter M Fischer, Tom De Nies, Erik Mannens e Rik Van de Walle. Information diffusion and provenance of interactions in twitter: Is it only about retweets? Em *Proceedings of the 25th International Conference Companion on World Wide Web*, páginas 113–114. International World Wide Web Conferences Steering Committee, 2016. 6, 20

[TL15] Chenhao Tan e Lillian Lee. All Who Wander : On the Prevalence and Characteristics of Multi-community Engagement. *WWW '15*, abs/1503.0, 2015. 24, 25

[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. 29

[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. 8

[vor] voreSnake. Dear reddit, a warning. https://www.reddit.com/r/reddit.com/comments/g5sdx/dear_reddit_a_warning/. 26

[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. 8, 23, 25, 33

[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. 33

[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. 6, 7

[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. 24

[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. 23